Automatic Evaluation of Machine Translation Based on Rate of Accomplishment of Sub-goals

Kiyotaka Uchimoto and **Katsunori Kotani** and **Yujie Zhang** and **Hitoshi Isahara**
National Institute of Information and Communications Technology
3-5, Hikari-dai, Seika-cho, Soraku-gun, Kyoto, 619-0289, Japan
{uchimoto,yujie,isahara}@nict.go.jp, kat@khn.nict.go.jp

Abstract

The quality of a sentence translated by a machine translation (MT) system is difficult to evaluate. We propose a method for automatically evaluating the quality of each translation. In general, when translating a given sentence, one or more conditions should be satisfied to maintain a high translation quality. In English-Japanese translation, for example, prepositions and infinitives must be appropriately translated. We show several procedures that enable evaluating the quality of a translated sentence more appropriately than using conventional methods. The first procedure is constructing a test set where the conditions are assigned to each test-set sentence in the form of yes/no questions. The second procedure is developing a system that determines an answer to each question. The third procedure is combining a measure based on the questions and conventional measures. We also present a method for automatically generating sub-goals in the form of yes/no questions and estimating the rate of accomplishment of the sub-goals. Promising results are shown.

1 Introduction

In machine translation (MT) research, appropriately evaluating the quality of MT results is an important issue. In recent years, many researchers have tried to automatically evaluate the quality of MT and improve the performance of automatic MT evaluations (Niessen et al., 2000; Akiba et al., 2001; Papineni et al., 2002; NIST, 2002; Leusch et al., 2003; Turian et al., 2003; Babych and Hartley, 2004; Lin and Och, 2004; Banerjee and Lavie, 2005; Gimeńez et al., 2005) because improving the performance of automatic MT evaluation is expected to enable us to use and improve MT systems efficiently. For example, Och reported that the quality of MT results was improved by using automatic MT evaluation measures for the parameter tuning of an MT system (Och, 2003). This report shows that the quality of MT results improves as the performance of automatic MT evaluation improves.

MT systems can be ranked if a set of MT results for each system and their reference translations are given. Usually, about 300 or more sentences are used to automatically rank MT systems (Koehn, 2004). However, the quality of a sentence translated by an MT system is difficult to evaluate. For example, the results of five MTs into Japanese of the sentence "The percentage of stomach cancer among the workers appears to be the highest for any asbestos workers." are shown in Table 1. A conventional automatic evaluation method ranks the fifth MT result first although its human subjective evaluation is the lowest. This is because conventional methods are based on the similarity between a translated sentence and its reference translation, and they give the translated sentence a high score when the two sentences are globally similar to each other in terms of lexical overlap. However, in the case of the above example,

Proceedings of NAACL HLT 2007, pages 33–40,
Rochester, NY, April 2007.

Table 1: Examples of conventional automatic evaluations.

Original sentence	The percentage of stomach cancer among the workers appears to be the highest for any asbestos workers.
Reference translation (in Japanese)	*roudousha no igan no wariai wa , asubesuto roudousha no tame ni saikou to naru youda .*

System	MT results	BLEU	NIST	Fluency	Adequacy
1	*roudousha no aida no igan no paasenteeji wa , donoyouna ishiwata roudousha no tame ni demo mottomo ookii youdearu .*	0.2111	2.1328	2	3
2	*roudousha no aida no igan no paasenteeji wa, arayuru asubesuto roudousha no tame ni mottomo takai youni omowa re masu .*	0.2572	2.1234	2	3
3	*roudousha no aida no igan no paasenteeji wa donna asubesuto no tame ni mo mottomo takai youni mie masu*	0	1.8094	1	2
4	*roudousha no aida no igan no paasenteeji wa nin'ino ishiwata ni wa mottomo takaku mie masu .*	0	1.5902	1	2
5	*roudousha no naka no igan no wariai wa donna asubesuto ni mo mottomo takai youni mieru .*	0.2692	2.2640	1	2

the most important thing to maintain a high translation quality is to correctly translate "for" into the target language, and it would be difficult to detect the importance just by comparing an MT result and its reference translations even if the number of reference translations is increased.

In general, when translating a given sentence, one or more conditions should be satisfied to maintain a high translation quality. In this paper, we show that constructing a test set where the conditions that are mainly established from a linguistic point of view are assigned to each test-set sentence in the form of yes/no questions, developing a system that determines an answer to each question, and combining a measure based on the questions and conventional measures enable the evaluation of the quality of a translated sentence more appropriately than using conventional methods. We also present a method for automatically generating sub-goals in the form of yes/no questions and estimating the rate of accomplishment of the sub-goals.

2 Test Set for Evaluating Machine Translation Quality

2.1 Test Set

Two main types of data are used for evaluating MT quality. One type of data is constructed by arbitrarily collecting sentence pairs in the source- and target-languages, and the other is constructed by intensively collecting sentence pairs that include linguistic phenomena that are difficult to automatically translate. Recently, MT evaluation campaigns such as the International Workshop on Spoken Language Translation [1], NIST Machine Translation Evaluation [2], and HTRDP Evaluation [3] were organized to support the improvement of MT techniques. The data used in the evaluation campaigns were arbitrarily collected from newspaper articles or travel conversation data for fair evaluation. They are classified as the former type of data mentioned above. On the other hand, the data provided by NTT (Ikehara et al., 1994) and that constructed by JEIDA (Isahara, 1995) are classified as the latter type. Almost all the data mentioned above consist of only parallel translations in two languages. Data with information for evaluating MT results, such as JEIDA's are rarely found. In this paper, we call data that consist of parallel translations collected for MT evaluation and that the information for MT evaluation is assigned to, a *test set*.

The most characteristic information assigned to the JEIDA test set is the yes/no question for assessing the translation results. For example, a yes/no question such as "Is 'for' translated into an expression representing a cause/reason such as 'de'?" (in Japanese) is assigned to a test-set sentence. We can evaluate MT results objectively by answering the question. An example of a test-set sample consisting of an ID, a source-language sample sentence, its reference translation, and a question is as follows.

[1] http://www.slt.atr.jp/IWSLT2006/
[2] http://www.nist.gov/speech/tests/mt/index.htm
[3] http://www.863data.org.cn/

ID	1.1.7.1.3-1
Sample sentence	The percentage of stomach cancer among the workers appears to be the highest for any asbestos workers.
Reference translation (in Japanese)	*roudousha no igan no wariai wa , asubesuto roudousha no tame ni saikou to naru youda .*
Question	Is "appear to" translated into an auxiliary verb such as "*youda*"?

The questions are classified mainly in terms of grammar, and the numbers to the left of the hyphenation of each ID such as 1.1.7.1.3 represent the categories of the questions. For example, the above question is related to catenative verbs.

The JEIDA test set consists of two parts, one for the evaluation of English-Japanese MT and the other for that of Japanese-English MT. We focused on the part for English-Japanese MT. This part consists of 769 sample sentences, each of which has a yes/no question.

The 769 sentences were translated by using five commercial MT systems to investigate the relationship between subjective evaluation based on yes/no questions and conventional subjective evaluation based on fluency and adequacy. The instruction for the subjective evaluation based on fluency and adequacy followed that given in the TIDES specification (TIDES, 2002). The subjective evaluation based on yes/no questions was done by manually answering each question for each translation. The subjective evaluation based on the yes/no questions was stable; namely, it was almost independent of the human subjects in our preliminary investigation. There were only two questions for which the answers generated inconsistency in the subjective evaluation when 1,500 question-answer pairs were randomly sampled and evaluated by two human subjects.

Then, we investigated the correlation between the two types of subjective evaluation. The correlation coefficients mentioned in this paper are statistically significant at the 1% or less significance level. The Spearman rank-order correlation coefficient is used in this paper. In the subjective evaluation based on yes/no questions, yes and no were numerically transformed into 1 and -1. For 3,845 translations obtained by using five MT systems, the correlation coefficients between the subjective evaluations based on yes/no questions and based on fluency and adequacy were 0.48 for fluency and 0.63 for adequacy. These results indicate that the two subjective evaluations have relatively strong correlations. The correlation is especially strong between the subjective evaluation based on yes/no questions and adequacy.

2.2 Expansion of JEIDA Test Set

Each sample sentence in the JEIDA test set has only one question. Therefore, in the subjective evaluation using the JEIDA test set, translation errors that do not involve the pre-assigned question are ignored even if they are serious. Therefore, translations that have serious errors that are not related to the question tend to be evaluated as being of high quality. To solve this problem, we expanded the test set by adding new questions about translations with the serious errors.

Sentences whose average grades were three or less for fluency and adequacy for the translation results of the five MT systems were selected for the expansion. Besides them, sentences whose average grades were more than three for fluency and adequacy for the translation results of the five MT systems were selected when a majority of evaluation results based on yes/no questions about the translations of the five MT systems were no. The number of selected sentences was 150. The expansion was manually performed using the following steps.

1. Serious translation errors are extracted from the MT results.
2. For each extracted error, questions strongly related to the error are searched for in the test set. If related questions are found, the same types of questions are generated for the selected sentence, and the same ID as that of the related question is assigned to each generated question. Otherwise, questions are newly generated, and a new ID is assigned to each generated question.
3. Each MT result is evaluated according to each added question.

Eventually, one or more questions were assigned to each selected sentence in the test set. Among the 150

Table 2: Expanded test-set samples.

Original	ID	1.1.7.1.3-1
	Sample sentence	The percentage of stomach cancer among the workers appears to be the highest for any asbestos workers.
	Reference translation (in Japanese)	*roudousha no igan no wariai wa , asubesuto roudousha no tame ni saikou to naru youda* .
	Question (Q-0)	Is "appear to" translated into an auxiliary verb such as "*youda*"?
Expanded	ID	1.1.6.1.3-5
	Translation error	"For" is not translated appropriately.
	Question-1 (Q-1)	Is "for" translated into an expression representing a cause/reason such as "*. . .de*"?
Expanded	ID	Additional-1
	Translation error	Some expressions are not translated.
	Question-2 (Q-2)	Are all English words translated into Japanese?

Table 3: Examples of subjective evaluations based on yes/no questions.

System	MT results	Answer Q-0	Q-1	Q-2	Fluency	Adequacy
1	*roudousha no aida no igan no paasenteeji wa , donoyouna ishiwata roudousha no tame ni demo mottomo ookii youdearu .*	Yes	No	Yes	2	3
2	*roudousha no aida no igan no paasenteeji wa, arayuru asubesuto roudousha no tame ni mottomo takai youni omowa re masu .*	Yes	Yes	Yes	2	3
3	*roudousha no aida no igan no paasenteeji wa donna asubesuto no tame ni mo mottomo takai youni mie masu*	Yes	No	No	1	2
4	*roudousha no aida no igan no paasenteeji wa nin'ino ishiwata ni wa mottomo takaku mie masu .*	Yes	No	No	1	2
5	*roudousha no naka no igan no wariai wa donna asubesuto ni mo mottomo takai youni mieru .*	Yes	No	No	1	2

selected sentences, questions were newly assigned to 103 sentences. The number of added questions was 148. The maximum number of questions added to a sentence was five. After expanding the test set, the correlation coefficients between the subjective evaluations based on yes/no questions and based on fluency and adequacy increased from 0.48 to 0.51 for fluency and from 0.63 to 0.66 for adequacy. The differences between the correlation coefficients obtained before and after the expansion are statistically significant at the 5% or less significance level for adequacy. These results indicate that the expansion of the test set significantly improves the correlation between the subjective evaluations based on yes/no questions and based on adequacy. When two or more questions were assigned to a test-set sentence, the subjective evaluation based on the questions was decided by the majority answer. The majority answers, yes and no, were numerically transformed into 1 and -1. Ties between yes and no were transformed into 0. Examples of added questions and the subjective evaluations based on the questions are shown in Tables 2 and 3.

3 Automatic Evaluation of Machine Translation Based on Rate of Accomplishment of Sub-goals

3.1 A New Measure for Evaluating Machine Translation Quality

The JEIDA test set was not designed for automatic evaluation but for human subjective evaluation. However, a measure for automatic MT evaluation that strongly correlates fluency and adequacy is likely to be established because the subjective evaluation based on yes/no questions has a relatively strong correlation with the subjective evaluation based on fluency and adequacy, as mentioned in Section 2. In this section, we describe a method for automatically evaluating MT quality by predicting an answer to each yes/no question and using those answers.

Hereafter, we assume that each yes/no question is defined as a sub-goal that a given translation should satisfy and that the sub-goal is accomplished if the answer to the corresponding yes/no question to the sub-goal is yes. We also assume that the sub-goal is unaccomplished if the answer is no. A new evaluation score, A, is defined based on a multiple lin-

Table 4: Examples of Patterns.

Sample sentence	She lived there by herself.
Question	Is "by herself" translated as "*hitori de*"?
Pattern	The answer is *yes* if the pattern [*hitori dake de*|*hitori kiri de* |*tandoku de*|*tanshin de*] is included in a translation. Otherwise, the answer is *no*.
Sample sentence	They speak English in New Zealand.
Question	The personal pronoun "they" is omitted in a translation like "*nyuujiilando de wa eigo wo hanasu*"?
Pattern	The answer is *yes* if the pattern [*karera wa*|*sore ra wa*] is not included in a translation. Otherwise, the answer is *no*.

ear regression model as follows using the rate of accomplishment of the sub-goals and the similarities between a given translation and its reference translation. The best-fitted line for the observed data is calculated by the method of least-squares (Draper and Smith, 1981).

$$A = \sum_{i=1}^{m} \lambda_{S_i} \times S_i + \sum_{j=1}^{n} (\lambda_{Q_j} \times Q_j + \lambda_{Q'_j} \times Q'_j) + \lambda_\epsilon \quad (1)$$

$$Q_j = \begin{cases} 1 : \text{if subgoal is accomplished} \\ 0 : \text{otherwise} \end{cases} \quad (2)$$

$$Q'_j = \begin{cases} 1 : \text{if subgoal is unaccomplished} \\ 0 : \text{otherwise} \end{cases} \quad (3)$$

Here, the term Q_j corresponds to the rate of accomplishment of the sub-goal having the i-th ID, and λ_{Q_j} is a weight for the rate of accomplishment. The term Q'_j corresponds to the rate of unaccomplishment of the sub-goal having the i-th ID, and $\lambda_{Q'_j}$ is a weight for the rate of unaccomplishment. The value n indicates the number of types of sub-goals. The term λ_ϵ is constant.

The term S_i indicates a similarity between a translated sentence and its reference translation, and λ_{S_i} is a weight for the similarity. Many methods for calculating the similarity have been proposed (Niessen et al., 2000; Akiba et al., 2001; Papineni et al., 2002; NIST, 2002; Leusch et al., 2003; Turian et al., 2003; Babych and Hartley, 2004; Lin and Och, 2004; Banerjee and Lavie, 2005; Gimeńez et al., 2005). In our research, 23 scores, namely BLEU (Papineni et al., 2002) with maximum n-gram lengths of 1, 2, 3, and 4, NIST (NIST, 2002) with maximum n-gram lengths of 1, 2, 3, 4, and 5, GTM (Turian et al., 2003) with exponents of 1.0, 2.0, and 3.0, METEOR (exact) (Banerjee and Lavie, 2005), WER (Niessen et al., 2000), PER (Leusch et al., 2003), and ROUGE (Lin, 2004) with n-gram lengths of 1, 2, 3, and 4 and 4 variants (LCS, S$*$, SU$*$, W-1.2), were used to calculate each similarity S_i. Therefore, the value of m in Eq. (1) was 23. Japanese word segmentation was performed by using JUMAN [4] in our experiments.

As you can see, the definition of our new measure is based on a combination of an evaluation measure focusing on local information and that focusing on global information.

3.2 Automatic Estimation of Rate of Accomplishment of Sub-goals

The rate of accomplishment of sub-goals is estimated by determining the answer to each question as yes or no. This section describes a method based on simple patterns for determining the answers.

An answer to each question is automatically determined by checking whether patterns are included in a translation or not. The patterns are constructed for each question. All of the patterns are expressed in *hiragana* characters. Before applying the patterns to a given translation, the translation is transformed into *hiragana* characters, and all punctuation is eliminated. The transformation to *hiragana* characters was performed by using JUMAN in our experiments.

Test-set sentences, the questions assigned to them, and the patterns constructed for the questions are shown in Table 4. In the patterns, the symbol "|" represents "OR".

3.3 Automatic Sub-goal Generation and Automatic Estimation of Rate of Accomplishment of Sub-goals

We found that expressions important for maintaining a high translation quality were often commonly

[4] http://www.kc.t.u-tokyo.ac.jp/nl-resource/juman.html

included in the reference translations for each test-set sentence. We also found that the expression was also related to the yes/no question assigned to the test-set sentence. Therefore, we automatically generate yes/no questions in the following steps.

1. For each test-set sentence, a set of words commonly appearing in the reference translations are extracted.

2. For each combination of n words in the set of words extracted in the first step, skip word n-grams commonly appearing in the reference translations in the same word order are selected as a set of common skip word n-grams.

3. For each test-set sentence, the sub-goal is defined as the yes/no question "Are all of the common skip word n-grams included in the translation?"

If no common skip word n-grams are found, the yes/no question is not generated. The answer to the yes/no question is determined to be yes if all of the common skip word n-grams are included in a translation. Otherwise, the answer is determined to be no.

This scheme assigns greater weight to important phrases that should be included in the translation to maintain a high translation quality. Our observation is that those important phrases are often common between human translations. A similar scheme was proposed by Babych and Hartley (Babych and Hartley, 2004) for BLEU. In their scheme, greater weight is assigned to components that are salient throughout the document. Therefore, their scheme focuses on global context while our scheme focuses on local context. We believe that the two schemes are complementary to each other.

4 Experiments and Discussion

In our experiments, the translation results of three MT systems and their subjective evaluation results were used as a development set for constructing the patterns described in Section 3.2 and for tuning the parameters λ_{S_i}, λ_{Q_j}, $\lambda_{Q_j'}$, and λ_ϵ in Eq. (1). The translations and evaluation results of the remaining two MT systems were used as an evaluation set for testing.

In the development set, each test-set sentence has at least one question, at least one reference translation, three MT results, and subjective evaluation results of the three MT results. The patterns for determining yes/no answers were manually constructed for the questions assigned to the 769 test-set sentences. There were 917 questions assigned to them. Among them, the patterns could be constructed for 898 questions assigned to 767 test-set sentences. The remaining 19 questions were skipped because making simple patterns as described in Section 3.2 was difficult; for example, one of the questions was "Is the whole sentence translated into one sentence?". The yes/no answer determination accuracies obtained by using the patterns are shown in Table 5.

Table 5: Results of yes/no answer determination.

Test set	Accuracy
Development	97.6% (2,629/2,694)
Evaluation	82.8% (1,487/1,796)

We investigated the correlation between the evaluation score, A in Eq. (1) and the subjective evaluations, fluency and adequacy, for the 769 test-set sentences. First, to maximize the correlation coefficients between the evaluation score, A, and the human subjective evaluations, fluency and adequacy, the optimal values of λ_{S_i}, λ_{Q_j}, $\lambda_{Q_j'}$, and λ_ϵ in Eq. (1) were investigated using the development set within a framework of multiple linear regression modeling (Draper and Smith, 1981). Then, the correlation coefficients were investigated by using the optimal value set. The results are shown in Table 6, 7, and 8. In these tables, "Conventional method" indicates the correlation coefficients obtained when A was calculated by using only similarities S_i. "Conventional method (combination)" is a combination of existing automatic evaluation methods from the literature. "Our method (automatic)" indicates the correlation coefficients obtained when the results of the automatic determination of yes/no answers were used to calculate Q_j and $Q_j^{'}$ in Eq. (1). For the 19 questions for which the patterns could not be constructed, Q_j was set at 0. "Our method (full automatic)" indicates the correlation coefficients obtained when the results of the automatic sub-goal generation and determination of rate of accomplish-

Table 6: Coefficients of correlation between evaluation score A and fluency/adequacy. (A reference translation is used to calculate S_i.)

Method	fluency		adequacy	
	Development set	Evaluation set	Development set	Evaluation set
Conventional method (WER)	0.43	0.48	0.42	0.48
Conventional method (combination)	0.52	0.51	0.49	0.47
Our method (automatic)	0.90$*$	0.59$*$	0.89$*$	0.62$*$
Our method (upper bound)	0.90$*$	0.62$*$	0.90$*$	0.68$*$

Table 7: Coefficients of correlation between evaluation score A and fluency/adequacy. (Three reference translations are used to calculate S_i.)

Method	fluency		adequacy	
	Development set	Evaluation set	Development set	Evaluation set
Conventional method (WER)	0.47	0.51	0.45	0.51
Conventional method (combination)	0.54	0.54	0.51	0.52
Our method (automatic)	0.90$*$	0.60$*$	0.90$*$	0.64$*$
Our method (full automatic)	0.85$*$	0.58	0.84$*$	0.60$*$
Our method (upper bound)	0.90$*$	0.62$*$	0.90$*$	0.69$*$

Table 8: Coefficients of correlation between evaluation score A and fluency/adequacy. (Five reference translations are used to calculate S_i.)

Method	fluency		adequacy	
	Development set	Evaluation set	Development set	Evaluation set
Conventional method (WER)	0.49	0.53	0.46	0.53
Conventional method (combination)	0.56	0.56	0.52	0.54
Our method (automatic)	0.90$*$	0.60	0.90$*$	0.63$*$
Our method (full automatic)	0.86$*$	0.59	0.85$*$	0.60$*$
Our method (upper bound)	0.91$*$	0.63$*$	0.90$*$	0.69$*$

In these tables, $*$ indicates significance at the 5% or less significance level.

ment of sub-goals were used to calculate Q_j and Q'_j in Eq. (1). Skip word trigrams, skip word bigrams, and skip word unigrams were used for generating the sub-goals according to our preliminary experiments. "Our method (upper bound)" indicates the correlation coefficients obtained when human judgments on the questions were used to calculate Q_j and Q'_j.

As shown in Table 6, 7, and 8, our methods significantly outperform the conventional methods from literature. Note that WER outperformed other individual measures like BLEU and NIST in our experiments, and the combination of existing automatic evaluation methods from the literature outperformed individual lexical similarity measures by themselves in almost all cases. The differences between the correlation coefficients obtained using our method and the conventional methods are statistically significant at the 5% or less significance level for fluency and adequacy, even if the number of reference translations increases, except in three cases shown in Table 7 and 8. This indicates that considering the rate of accomplishment of sub-goals to automatically evaluate the quality of each translation is useful, especially when the number of reference translations is small.

The differences between the correlation coefficients obtained using two automatic methods are not significant. These results indicate that we can reduce the development cost for constructing sub-goals. However, there are still significant gaps between the correlation coefficients obtained using a fully automatic method and upper bounds. These gaps indicate that we need further improvement in automatic sub-goal generation and automatic estimation of rate of accomplishment of sub-goals, which is our future work.

Human judgments of adequacy and fluency are known to be noisy, with varying levels of intercoder agreement. Recent work has tended to apply cross-judge normalization to address this issue (Blatz et al., 2003). We would like to evaluate against the normalized data in the future.

5 Conclusion and Future Work

We demonstrated that the quality of a translated sentence can be evaluated more appropriately than by using conventional methods. That was demonstrated by constructing a test set where the conditions that should be satisfied to maintain a high translation quality are assigned to each test-set sentence in the form of a question, by developing a system that determines an answer to each question, and by combining a measure based on the questions and conventional measures. We also presented a method for automatically generating sub-goals in the form of yes/no questions and estimating the rate of accomplishment of the sub-goals. Promising results were obtained.

In the near future, we would like to expand the test set to improve the upper bound obtained by our method. We are also planning to expand the method and improve the accuracy of the automatic sub-goal generation and determination of the rate of accomplishment of sub-goals. The sub-goals of a given sentence should be generated by considering the complexity of the sentence and the alignment information between the original source-language sentence and its translation. Further advanced generation and estimation would give us information about the erroneous parts of MT results and their quality. We believe that future research would allow us to develop high-quality MT systems by tuning the system parameters based on the automatic MT evaluation measures.

Acknowledgments

The guideline for expanding the test set is based on that constructed by the Technical Research Committee of the AAMT (Asia-Pacific Association for Machine Translation) The authors would like to thank the committee members, especially, Mr. Kentaro Ogura, Ms. Miwako Shimazu, Mr. Tatsuya Sukehiro, Mr. Masaru Fuji, and Ms. Yoshiko Matsukawa for their cooperation. This research is partially supported by special coordination funds for promoting science and technology.

References

Yasuhiro Akiba, Kenji Imamura, and Eiichiro Sumita. 2001. Using Multiple Edit Distances to Automatically Rank Machine Translation Output. In *Proceedings of the MT Summit VIII*, pages 15–20.

Bogdan Babych and Anthony Hartley. 2004. Extending the BLEU MT Evaluation Method with Frequency Weightings. In *Proceedings of the 42nd ACL*, pages 622–629.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, pages 65–72.

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing. 2003. Confidence Estimation for Machine Translation. Technical report, Center for Language and Speech Processing, Johns Hopkins University. Summer Workshop Final Report.

Norman R. Draper and Harry Smith. 1981. *Applied Regression Analysis. 2nd edition*. Wiley.

Jesuś Gimeńez, Enrique Amigo, and Chiori Hori. 2005. Machine translation evaluation inside qarla. In *Proceedings of the IWSLT'05*.

Satoru Ikehara, Satoshi Shirai, and Kentaro Ogura. 1994. Criteria for Evaluating the Linguistic Quality of Japanese to English Machine Translations. *Transactions of the JSAI*, 9(4):569–579. (in Japanese).

Hitoshi Isahara. 1995. JEIDA's Test-Sets for Quality Evaluation of MT Systems – Technical Evaluation from the Developer's Point of View.

Philipp Koehn. 2004. Statistical Significance Tests for Machine Translation Evaluation. In *Proceedings of the 2004 Conference on EMNLP*, pages 388–395.

Gregor Leusch, Nicola Ueffing, and Hermann Ney. 2003. A Novel String-to-String Distance Measure with Applications to Machine Translation Evaluation. In *Proceedings of the MT Summit IX*, pages 240–247.

Chin-Yew Lin and Franz Josef Och. 2004. ORANGE: a Method for Evaluating Automatic Evaluation Metrics for Machine Translation. In *Proceedings of the 20th COLING*, pages 501–507.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Proceedings of the Workshop on Text Summarization Branches Out*, pages 74–81.

Sonja Niessen, Franz Josef Och, Gregor Leusch, and Hermann Ney. 2000. An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. In *Proceedings of the LREC 2000*, pages 39–45.

NIST. 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. Technical report, NIST.

Franz Josef Och. 2003. Minimum Error Training in Statistical Machine Translation. In *Proceedings of the 41st ACL*, pages 160–167.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th ACL*, pages 311–318.

TIDES. 2002. Linguistic Data Annotation Specification: Assessment of Fluency and Adequacy in Arabic-English and Chinese-English Translations. http://www.ldc.upenn.edu/Projects/TIDES/Translation/TransAssess02.pdf.

Joseph P. Turian, Luke Shen, and I. Dan Melamed. 2003. Evaluation of Machine Translation and its Evaluation. In *Proceedings of the MT Summit IX*, pages 386–393.