The Effect of Adding Rules into the Rule-based MT System

ZHU Jiang, WANG Haifeng
Toshiba (China) Research and Development Center
5/F., Tower W2, Oriental Plaza
No.1, East Chang An Ave., Dong Cheng District
Beijing, 100738, China
{zhujiang, wanghaifeng}@rdc.toshiba.com.cn

Abstract

This paper investigates the relationship between the amount of the rules and the performance of the rule-based machine translation system. We keep adding more rules into the system and observe successive changes of the translation quality. Evaluations on translation quality reveal that the more the rules, the better the translation quality. A linear regression analysis shows that a positive linear relationship exists between the translation quality and the amount of the rules. We use this linear model to make prediction and test the prediction with newly developed rules. Experimental results indicate that the linear model effectively predicts the possible performance that the rule-based machine translation system may achieve with more rules added.

1 Introduction

Rule-based approach is one of the most important approaches in the field of machine translation. It has been widely used to develop practical machine translation systems. Previous studies pertaining to rule-based approach include theoretical studies, or system implementation, etc. (Hutchins and Somers, 1992). However, little work has been done to quantitatively examine the effect of adding rules to the rule-based machine translation system. In fact, analysis of the relationship between the amount of the rules and the performance of the machine translation system could help to estimate the possible performance that the system may achieve in the future with more rules added to the system.

In this paper, we investigate the effect of adding rules to the rule-based machine translation system. We measure the performance of the system by evaluating the translation quality of the system. We keep adding more rules into the machine translation system and observe successive changes of the translation quality. Evaluation results reveal that the more the rules, the better the translation quality. Statistical analysis shows there is a linear relationship between the translation quality and the amount of the rules. Further experiment tests the linear model and predicts the possible performance that the rule-based machine translation system may achieve in the future with more rules added.

The motivation for this work arose from the questions we met when developing the rule-based machine translation system. We have an English-to-Chinese machine translation (ECMT) system that is under development. We keep adding rules to the rule base and periodically perform evaluations on the system to track the progress of the development. Questions were raised when we observed the durative improvement of the system. We were interested in answering questions such as how fast the performance will improve? Is there a model that can help us to predict the improvement in the near future? Can we estimate the cost of achieving the improvement so that we can make our development plan more reasonably? What is the potential of the system? We hope our questions could be answered by careful analysis of the relationship between the performance and the amount of the rules.

In this paper, we measure the performance of the system by translation quality. Several kinds of approaches (White et al., 1994; Nießen et al., 2000; Akiba, et al., 2001; Turian et al., 2003) can be used to evaluate translation quality. In our experiment, we adopt n-gram co-occurrence based automatic evaluation method that was first proposed by (Papineni et al., 2002). We use the NIST scoring toolkit[1], which is based on the modified n-gram co-occurrence method (Doddington 2002) and has been used in the machine translation evaluations organized by DARPA TIDES[2]. As for the rule-based machine translation system, we take our ECMT system as a case study. All the experiment

[1] The NIST scoring toolkit could be downloaded from: http://www.nist.gov/speech/tests/mt/resources/scoring.htm

[2] Related information about the MT evaluation could be found at: http://www.nist.gov/speech/tests/mt/index.htm

results reported in this paper are based on this ECMT system.

The rest of this paper is organized as follows. In section 2, we give a brief introduction to our rule-based ECMT system and the semantic rules used in the system. In section 3 we describe our methodology. Section 4 reports the experimental results, presents the linear regression analysis, tests the effectiveness of the linear model and predicts the improvement of the ECMT system. We further discuss the experiment results in section 5. Finally we present our conclusion and the future work in section 6.

2 The Rule-based ECMT System

We have been developing a rule-based ECMT system. The basis of the system is semantic transfer, for which a grammatical model called 'Lexical Transition Network Grammar' was developed. This model is powerful for semantic analysis, which is performed by lexical grammars attached to each item in the lexicon (Amano et al., 1989).

We take five steps in the translation flow: (1) dictionary and morphological analysis; (2) syntactic analysis; (3) semantic analysis; (4) structural transfer; (5) generation.

For the sake of simplicity, in this research we only consider semantic rules of the system and keep the other parts of the system untouched. That is, we only study the effect of adding semantic rules into the rule base of the system[3].

The semantic rules are attached to words. They are used not only for semantic analysis but also for transferring the source language to the target language. They have following functions:

Lexical transfer and selection of translation. The rules transfer words of the source sentence into the target language. If there are alternative translations, the lexical rules select the proper translation according to the context in the sentence. For example, for the word "bank", it has a rule that set "岸" as the translation when it means the land along the side of a river or lake. It also has a rule that set "银行" as the translation when it means a business that keeps and lends money and provides other financial services.

Processing of idioms. For a sequence of words that must be treated as a single unit for translation purposes, semantic analysis interprets these words in the syntactic structure as an idiom using lexical rules attached to the headword of the idiom. For example, a rule attached to the word "catch" interprets "catch up with" as an idiom and sets the translation "赶上" for it.

Lexical rules with structural transfer. This is to perform structure conversion during English to Chinese translation. For example, the following rule transfers the structure of "a group of":

group(***det***_a ***npp***_of(***np***_student))

➔ student(***class***_group(***det***_a))

It sets "group" as a classifier of the noun "student" in Chinese.

3 Methodology

3.1 Basic Rules vs. Advanced Rules

The rule base of the ECMT system contains a wide variety of semantic rules. Because the semantic rules are attached to words, the effect of the rule varies. Rules concerning the common usage of words have greater impact on translation than rules concerning the special usage of words. Similarly, rules concerning the basic sentence patterns usually make broader influence on translation. Hence, based on the relative importance of the rules, we classify the rules into two categories:

Basic Rules, which correlate to basic usages or structures;

Advanced Rules, which correlate to complicated usages or structures.

To make the classification criteria applicable, we define ***basic rules*** especially as the rules that cover the vocabulary and grammar of the middle-school English textbook used in China. It is reasonable because the words and grammar of the middle-school English textbook are assumed to be the basic knowledge a person should comprehend when he/she starts to learn English. Correspondingly, we define ***advanced rules*** as the rules regarding a variety of linguistic phenomena, which include comprehensive usage of words and some complicated structures beyond the middle-school English textbook.

According to above definitions, basic rules should have higher priority than advanced rules when making the development plan. So when we develop rules for the system, we started with developing basic rules. Afterwards, we developed advanced rules.

3.2 Random Hypothesis

Once the development of basic rules was done, the system could produce rough translation for a given input sentence. Though there might be modifications on the basic rules to complete them, the overall amount of basic rules will not change a lot during the development procedure. There are about three thousand basic rules, accounting for

[3] In the following part of this paper, rule base refers to semantic rule base; rule refers to semantic rule.

about one tenth of the current rule base of the system.

However, the number of advanced rules continues to rise with the development of the system. To improve the translation performance of the system, lots of work is required to develop advanced rules. Because advanced rules cover such a wide variety of linguistic phenomena, the development order tends to vary with different development plans. To create a general model for machine translation systems that adopt different development plans, we pose the following hypothesis:

The improvement of the performance is independent of the arrangement of the advanced rules. -- H_1

The hypothesis will be verified in section 4.4.

Based on this hypothesis, we randomly split the advanced rules into segments of equal size.

3.3 Modeling and Prediction

To observe the change of the performance with different amount of rules, we first remove all the advanced rules from the rule base and get a rule base containing only basic rules. We call this rule base the *Basic Rule Base*. With the Basic Rule Base the system produces rough translations for the input sentences.

Afterwards, we put back the segments of advanced rules into the rule base, one at a time, and measure performance of the system, until the rule base is restored. The performance of the system is evaluated by the NIST scoring tool. Thus we obtain a sample of scores corresponding to rule bases of different sizes. We conduct a regression analysis on the sample and build a regression model that describes the relationship between the performance and the amount of the rules.

Finally, we use the model to make prediction for the improvement of the system. We calculate the prediction interval based on the model, with a given probability. Then we introduce another set of newly developed rules to verify the prediction. Instead of adding the new rules into the rule base randomly, we add them in order of the development sequence. We obtain a new sample of scores with the new rules added and check if the sample of scores falls into the prediction interval.

4 Experiments

4.1 Testing Set

We conduct experiments with a 1000-sentence testing set. The 1000 testing sentences are randomly selected from our English-Chinese bilingual corpus that contains about 400,000 English-Chinese bilingual sentence pairs in general domain. None of the sentences contained in this bilingual corpus has been used to tune our ECMT system.

4.2 Automatic MT Evaluation

To evaluate the translation quality of the ECMT system with the NIST scoring tool, a set of reference translations is required besides the source sentences and the output translations produced by the system. The source set, reference set and output translation set are annotated according to predefined format. In particular, Chinese word segmentation is performed on both the output translation set and the reference set.

4.3 Rule Base

We use the advanced rules developed before Sept. 2004 to build the model. Advanced rules developed from Sept. 2004 to Dec. 2004 will be used to test the prediction of the model. We call this part of advanced rules testing rules.

The amount and the proportion of each kind of rules are shown in table 1.

Rule Type	Basic	Advanced	Testing
Amount	3263	19800	5400
Proportion	11%	70%	19%

Table 1. Amount and Proportion of Rules

To build the model, we randomly split the advanced rules into 66 segments, each containing 300 advanced rules.

We have following definitions:

$RBS = \{RB_i | i = 0,1,...,66\}$

$RB_0 = BRB$

$RB_i = RB_{i-1} + \Delta_i, i = 1,...,66$

RBS represents Rule Base Set;

RB represents Rule Base;

BRB represents Basic Rule Base;

Δ represents a random segment that contains 300 advanced rules.

4.4 Effect of Rule Base Size

Firstly we run the ECMT system with RB_0 (BRB) to translate the testing set and get the baseline translation output. Then for each $RB_{i(i=1,...,66)}$, we run the ECMT system based on it to translate the testing set and obtain 66 translation outputs. Lastly, we automatically evaluate all the translation outputs with the NIST scoring tool. Thus, we obtain the scores of translations based on $RB_{i(i=0,...,66)}$.

Both NIST score and BLEU score are reported by NIST scoring tool. The BLEU scores show a behavior very similar to that of the NIST scores. So we only report the results and conclusions based on the NIST scores in this paper and omit the BLEU ones.

From the result, we observe an approximate linear increase of the score by increment of advanced rules. We pose the following hypothesis based on the observation:

The score has a linear progressive increase with the amount of rules. -- H_2

To test H_2, we introduce the linear regression model:

$$Score(x) = a + bx \tag{1}$$

x denotes the sequence number of RB, $Score(x)$ is the NIST score of the system based on rule base RB_x, a is the intercept of the line and b is the slope of the line. The estimates of a and b are calculated as follows:

$$\hat{b} = \frac{n\sum x_i Score(x_i) - \sum x_i \sum Score(x_i)}{n\sum x_i^2 - (\sum x_i)^2} \tag{2}$$

$$\hat{a} = \frac{\sum Score(x_i)}{n} - \hat{b}\frac{\sum x_i}{n} \tag{3}$$

Where $i = 0,1,2,...,n-1, \quad n = 67$

According to our results, we have

$\hat{a} = 4.8064$, $\hat{b} = 0.0098$

As a result, we obtain the regression line:

$$\hat{S}core(x) = 4.8064 + 0.0098x$$

$\hat{S}core(x)$ is the estimate of $Score(x)$.

The scatter diagram for the scoring results and the regression line fitted to the data are shown in figure 1.

The statistic **correlation coefficient, *r*,** is used to measure how well the regression line describes the data. It can be calculated by

$$|r| = \sqrt{1 - \frac{Q}{L_{ss}}} \tag{4}$$

Where

$$Q = \sum_{i=0}^{n-1} (Score(x_i) - \hat{S}core(x_i))^2 \tag{5}$$

$$L_{ss} = \sum_{i=0}^{n-1} (Score\ (x_i) - \bar{S}core\ (x))^2 = \sum_{i=0}^{n-1} Score\ (x_i)^2 - n\bar{S}core\ (x)^2 \tag{6}$$

$$\bar{S}core(x) = \frac{\sum_{i=0}^{n-1} Score(x_i)}{n} \tag{7}$$

$n = 67$

If r is near 1, it indicates that the response variable and the independent variable have a strong linear association. If r is near zero, it suggests that the model is ineffective.

Based on equations (4), (5), (6) and (7), we get $r = 0.9857$, which indicates a strong linear correlation between x and $Score(x)$. r^2, the coefficient of determination, is frequently interpreted as the fraction of the variability explained by the independent variable x. Here we have $r^2 = 0.9716$, which means that 97.16% of the total variation in $Score(x)$ can be explained

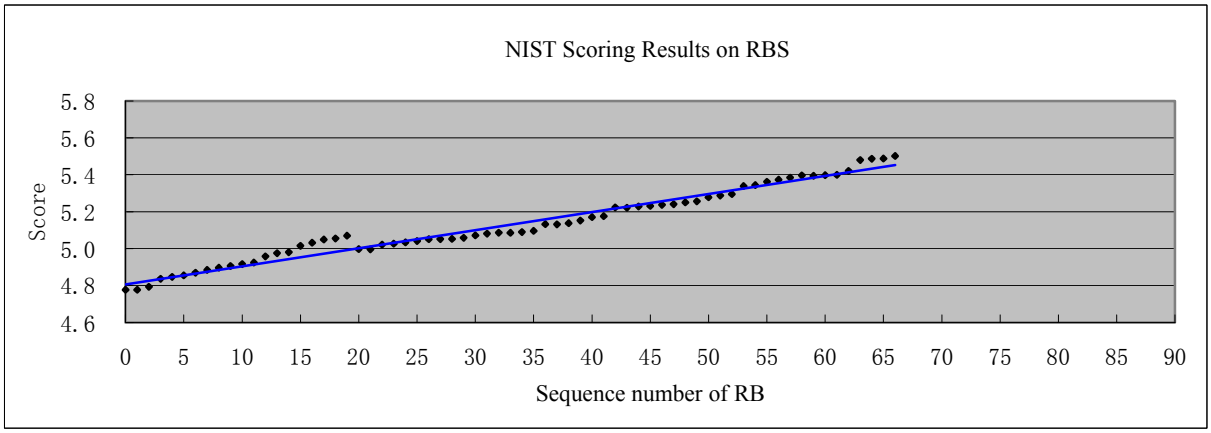

Figure 1. NIST Scoring Results on *RBS*

by the linear relationship between x and $Score(x)$ (as described by the regression model). So the regression is considered a good fit.

Now let us see the hypothesis H_1: ***The improvement of the performance is independent of the arrangement of the advanced rules***.

We generate another rule base set RBS' by repeating the random segmentation and restoring procedure. We have:

$$RBS' = \{RB_i' | i = 0,1,...,66\}$$

$$RB_0' = BRB$$

$$RB_i' = RB_{i-1}' + \Delta_i, i = 1,...,66$$

Considering the randomness, RB_i' is different from RB_i for i from 1 to 65. But RB_0' and RB_{66}' are identical with RB_0 and RB_{66}. Then we score the translation outputs produced by the system based on each $RB_{i(i=0,...,66)}'$ and get the following results as shown in figure 2.

According to equations (4), (5), (6) and (7), we get $r' = 0.9917$ and $r^2 = 0.9834$. So the relationship between the score and RB_i' could also be positively described by the regression line.

From the above discussion, we can reach the following conclusions:

(1) At current development stage, the improvement of the performance of the system has a linear relationship with the increase of advanced rules.

(2) The improvement of the performance of the system is independent of the development order of the advanced rules.

4.5 Prediction Experiment

We are also interested in using this model to predict the future performance of the system. We calculate the prediction interval for the model as follows.

For a given x, the $100(1-\alpha)\%$ prediction interval for $Score(x)$ is:

$$(\hat{S}core(x) - Z_{\alpha/2} S_\delta,\ \hat{S}core(x) + Z_{\alpha/2} S_\delta)$$

Where

$$S_\delta = \sqrt{\frac{\sum_{i=0}^{n-1} (Score(x_i) - \hat{S}core(x_i))^2}{n-2}} \qquad (8)$$

is the residual variance of $Score(x)$, which is used to indicate the error of $Score(x)$ that deviates from the regression line.

$Z_{\alpha/2}$ is the upper $100\frac{\alpha}{2}$ percentage point of the standard normal distribution.

This interval describes the area in which a new observation can be expected to fall with a certain probability.

We obtain the 99% prediction interval:

$$(\hat{S}core(x) - 0.0847,\ \hat{S}core(x) + 0.0847)$$

which is the area between the two lines as shown in figure 3.

Thus, for a given rule base of particular size, for example, a rule base contains 20100 advanced rules (that is to add 300 more advanced rules to RB_{66}), the NSIT score will be between 5.3783 and 5.5477 with 99% probability.

We use the testing rules to verify the prediction. Instead of randomly splitting the testing rules into segments, we sort the testing rules in order of

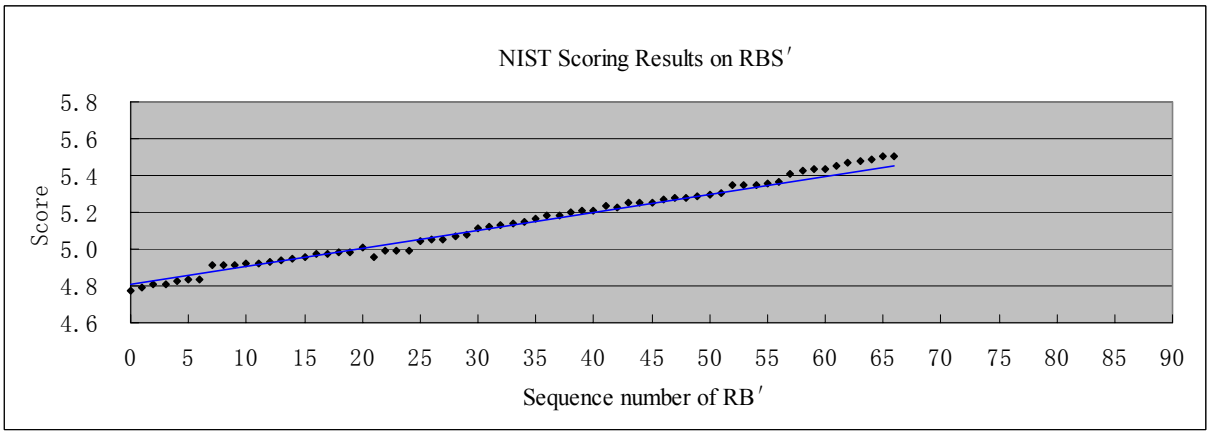

Figure 2. NIST Scoring Results on RBS' Comparing with the Regression Line

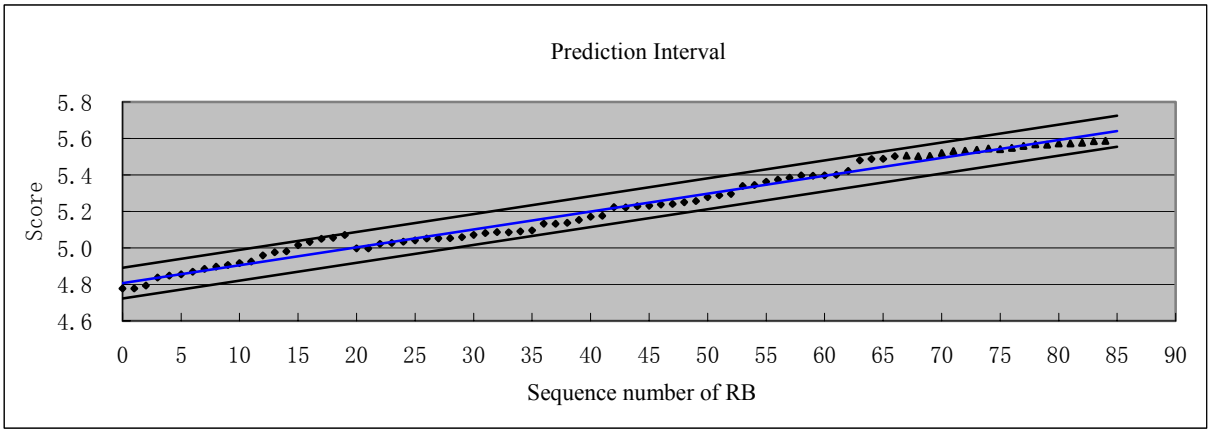

Figure 3. Predicting with the Model

development time and then equally divide them into 18 segments. Each segment contains 300 testing rules. We put these segments into the rule base orderly, one at each time, and measure the translation quality of the system, until all the segments are put into the rule base. The NIST scores are shown in figure 3 (the triangles). All of them fall into the prediction interval. From this we conclude that the results of our experiment with testing rules tend to support the model we obtained.

Now let us go back to the questions we presented in section 1. The improvement of the performance could be quantitatively predicted with the regression model and the prediction interval. So, with the estimation on workload and cost of developing a certain number of rules, we could estimate the investment and development period to achieve a prospective performance. This is of great value for drawing up a development plan.

5 Discussion

5.1 Potential of Rule-based Method

At current development stage, the improvement of the performance of the system has a linear relationship with the increase of advanced rules. It indicates that our system still has potential to be improved. Does it suggest that the rule-based method still has potential to improve the performance of the machine translation system? To answer this question, we compare our system with other rule-based ECMT systems. Table 2 shows the evaluation result on our ECMT system and six commercial ECMT systems using the same testing set. The six commercial ECMT systems are the mainstream ECMT systems available on Chinese market. All of them are rule-based machine translation systems. Evaluation shows that our system outperforms all the six commercial ECMT systems. Since the model shows that our system has the potential to be improved, it suggests, to a certain extent, that the rule-based method still has the potential to improve the performance of the machine translation system.

System	NIST Score
Ours	5.5860
Sys. 1	5.4850
Sys. 2	5.3283
Sys. 3	5.2555
Sys. 4	5.1046
Sys. 5	4.9254
Sys. 6	4.6599

Table 2. Evaluation on ECMT Systems

5.2 Interrelation Between Rules

In figure 1, a sudden drop is observed between $Score(RB_{19})$ and $Score(RB_{20})$. This is caused by the interrelation between rules. For example, we have the following three rules:

Rule A relating to the usage of "give sth. to sb."

Rule B relating to the usage of "give away"

Rule C relating to the usage of "give away sth. to sb."

Rule A is a basic rule. It translates the phrase "give sth. to sb." into "把…给…".

Rule B is an advanced rule. It translates the phrase "give away" into different Chinese translations such as "泄露", "暴露", "出卖", etc, according to different contexts. It also sets a default translation "送掉" for "give away".

Rule C is an advanced rule. It translates the phrase "give away sth. to sb." into "把…送给…".

Considering the following sentence:

I give away my pans to a friend.

我把我的盘子送给一位朋友。

The correct rule for translating this sentence is Rule C. However, before Rule B and Rule C are added to the rule base, the system translates the sentence following Rule A and gets the translation:

Tran A: 我**把**我的盘子**离开给**一位朋友。

This is an acceptable translation except the redundant translation of "away" (离开).

After we add Rule B to the rule base, the system translates the sentence according to Rule B and we get the translation:

Tran B: 我到一位朋友送掉我的盘子。

It is a worse translation than Tran A. So a sudden drop of translation quality is observed.

Once Rule C is added to the rule base, the system translates the sentence according to Rule C and we get a perfect translation:

Tran C: 我把我的盘子送给一位朋友。

The interrelationship between rules is a problem that always occurs in the rule-based machine translation system. In our rule-based ECMT system, the structure of the rule base is carefully designed and the rules are fine-grained in order to alleviate the problem of interrelationship between rules. The hierarchical semantic rule base contains rules at three different levels: post-parsing, pre-transfer and transfer. Rules of the same level are sorted according to their coverage. However, the interrelationship still cannot be completely avoided.

6 Conclusions and Future Work

In this paper we propose a linear regression model to describe the relationship between the rule base increment and the performance of the rule-based machine translation system. The model is proved to be effective and could positively represent the relationship and make prediction. Our main contribution is that, we present a novel method to quantitatively describe and predict the improvement of the rule-based machine translation system, which is significant for researchers and developers to make better development plans. Furthermore, the trend line described by the model shows that the rule-based method still has great potential to improve the translation performance of the machine translation system.

In future work, we will keep tracking the improvement of the system and observe the relationship between the improvement of the performance and the increment of rules. Besides analyzing the influence of semantic rules, we plan to analyze the relationship between the performance of the system and other kinds of rules in the future. In addition, we will consider applying our approach to machine translation systems based on other methods.

Acknowledgements

The authors would like to thank He Lei, Ge Naisheng and Guo Yuqing for their help on the preparation and study of this work. We also wish to thank the anonymous reviewers for their valuable comments.

References

Yasuhiro Akiba, Kenji Imamura and Eiichiro Sumita. 2001. Using Multiple Edit Distances to Automatically Rank Machine Translation Output. *In Proceedings of the 8th Machine Translation Summit (MT Summit VIII),* pages 15–20, Santiago de Compostela, Spain.

Shin-ya Amano, Hideki Hirakawa, Hirosysu Nogami, and Akira Kumano. 1989. The Toshiba Machine Translation system. *Future Computing System,* 2(3):227-246.

George Doddington. 2002. Automatic Evaluation of Machine Translation Quality Using N-gram Co-occurrence Statistics. *In Proceeding of the Second International Conference on Human Language Technology (HLT-2002),* pages 138-145, San Diego, CA, USA.

W. John Hutchins and Harold L. Somers. 1992. *An introduction to machine translation*. Academic Press, London, UK.

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 2nd International Conference on Language Resources and Evaluation (LREC-2000),* pages 39-45, Athens, Greece.

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 Annual Meeting of the Association for Computational Linguistics (ACL-2002),* pages 311-318, Philadelphia, PA, USA.

Joseph P. Turian, Luke Shen, and I. Dan Melamed. 2003. Evaluation of Machine Translation and its Evaluation. *In Proceedings of the 9th Machine Translation Summit (MT Summit IX),* pages 23-28, New Orleans, LA, USA.

John S. White, Theresa O'Connell and Francis O'Mara. 1994. The ARPA MT Evaluation Methodologies: Evolution, Lessons and Future Approaches. *In Proceedings of the 1st Conference of the Association for Machine Translation in the Americas (AMTA-1994),* pages 193-205, Columbia, MD, USA.