Feedback Cleaning of Machine Translation Rules Using Automatic Evaluation

Kenji Imamura, Eiichiro Sumita
ATR Spoken Language Translation
Research Laboratories
Seika-cho, Soraku-gun, Kyoto, Japan
{kenji.imamura,eiichiro.sumita}@atr.co.jp

Yuji Matsumoto
Nara Institute of
Science and Technology
Ikoma-shi, Nara, Japan
matsu@is.aist-nara.ac.jp

Abstract

When rules of transfer-based machine translation (MT) are automatically acquired from bilingual corpora, incorrect/redundant rules are generated due to acquisition errors or translation variety in the corpora. As a new countermeasure to this problem, we propose a feedback cleaning method using automatic evaluation of MT quality, which removes incorrect/redundant rules as a way to increase the evaluation score. BLEU is utilized for the automatic evaluation. The hill-climbing algorithm, which involves features of this task, is applied to searching for the optimal combination of rules. Our experiments show that the MT quality improves by 10% in test sentences according to a subjective evaluation. This is considerable improvement over previous methods.

1 Introduction

Along with the efforts made in accumulating bilingual corpora for many language pairs, quite a few machine translation (MT) systems that automatically acquire their knowledge from corpora have been proposed. However, knowledge for transfer-based MT acquired from corpora contains many incorrect/redundant rules due to acquisition errors or translation variety in the corpora. Such rules conflict with other existing rules and cause implausible MT results or increase ambiguity. If incorrect rules could be avoided, MT quality would necessarily improve.

There are two approaches to overcoming incorrect/redundant rules:

- Selecting appropriate rules in a disambiguation process during the translation (on-line processing, (Meyers et al., 2000)).
- Cleaning incorrect/redundant rules after automatic acquisition (off-line processing, (Menezes and Richardson, 2001; Imamura, 2002)).

We employ the second approach in this paper. The cutoff by frequency (Menezes and Richardson, 2001) and the hypothesis test (Imamura, 2002) have been applied to clean the rules. The cutoff by frequency can slightly improve MT quality, but the improvement is still insufficient from the viewpoint of the large number of redundant rules. The hypothesis test requires very large corpora in order to obtain a sufficient number of rules that are statistically confident.

Another current topic of machine translation is automatic evaluation of MT quality (Papineni et al., 2002; Yasuda et al., 2001; Akiba et al., 2001). These methods aim to replace subjective evaluation in order to speed up the development cycle of MT systems. However, they can be utilized not only as developers' aids but also for automatic tuning of MT systems (Su et al., 1992).

We propose **feedback cleaning** that utilizes an automatic evaluation for removing incorrect/redundant translation rules as a tuning method

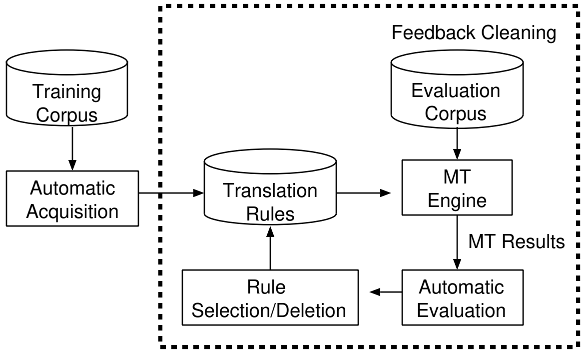

Figure 1: Structure of Feedback Cleaning

(Figure 1). Our method evaluates the contribution of each rule to the MT results and removes inappropriate rules as a way to increase the evaluation scores. Since the automatic evaluation correlates with a subjective evaluation, MT quality will improve after cleaning.

Our method only evaluates MT results and does not consider various conditions of the MT engine, such as parameters, interference in dictionaries, disambiguation methods, and so on. Even if an MT engine avoids incorrect/redundant rules by on-line processing, errors inevitably remain. Our method cleans the rules in advance by only focusing on the remaining errors. Thus, our method complements on-line processing and adapts translation rules to the given conditions of the MT engine.

2 MT System and Problems of Automatic Acquisition

2.1 MT Engine

We use the Hierarchical Phrase Alignment-based Translator (HPAT) (Imamura, 2002) as a transfer-based MT system. The most important knowledge in HPAT is transfer rules, which define the correspondences between source and target language expressions. An example of English-to-Japanese transfer rules is shown in Figure 2. The transfer rules are regarded as a synchronized context-free grammar.

When the system translates an input sentence, the sentence is first parsed by using source patterns of the transfer rules. Next, a tree structure of the target language is generated by mapping the source patterns to the corresponding target patterns. When non-terminal symbols remain in the target tree, target words are inserted by referring to a translation dictionary.

Ambiguities, which occur during parsing or mapping, are resolved by selecting the rules that minimize the semantic distance between the input words and source examples (real examples in the training corpus) of the transfer rules (Furuse and Iida, 1994). For instance, when the input phrase "*leave at 11 a.m.*" is translated into Japanese, Rule 2 in Figure 2 is selected because the semantic distance from the source example (*arrive*, *p.m.*) is the shortest to the head words of the input phrase (*leave*, *a.m.*).

2.2 Problems of Automatic Acquisition

HPAT automatically acquires its transfer rules from parallel corpora by using Hierarchical Phrase Alignment (Imamura, 2001). However, the rule set contains many incorrect/redundant rules. The reasons for this problem are roughly classified as follows.

- Errors in automatic rule acquisition
- Translation variety in corpora
  - The acquisition process cannot generalize the rules because bilingual sentences depend on the context or the situation.
  - Corpora contain multiple (paraphrasable) translations of the same source expression.

In the experiment of Imamura (2002), about 92,000 transfer rules were acquired from about 120,000 bilingual sentences [1]. Most of these rules are low-frequency. They reported that MT quality slightly improved, even though the low-frequency rules were removed to a level of about 1/9 the previous number. However, since some of them, such as idiomatic rules, are necessary for translation, MT quality cannot be dramatically improved by only removing low-frequency rules.

3 Automatic Evaluation of MT Quality

We utilize BLEU (Papineni et al., 2002) for the automatic evaluation of MT quality in this paper.

BLEU measures the similarity between MT results and translation results made by humans (called

[1] In this paper, the number of rules denotes the number of unique pairs of source patterns and target patterns.

Rule No.	Syn. Cat.	Source Pattern		Target Pattern	Source Example
1	VP	$X_{\mathtt{VP}}$ *at* $Y_{\mathtt{NP}}$	⇒	Y' *de* X'	((*present*, *conference*) ...)
2	VP	$X_{\mathtt{VP}}$ *at* $Y_{\mathtt{NP}}$	⇒	Y' *ni* X'	((*stay*, *hotel*), (*arrive*, *p.m*) ...)
3	VP	$X_{\mathtt{VP}}$ *at* $Y_{\mathtt{NP}}$	⇒	Y' *wo* X'	((*look*, *it*) ...)
4	NP	$X_{\mathtt{NP}}$ *at* $Y_{\mathtt{NP}}$	⇒	Y' *no* X'	((*man*, *front desk*) ...)

Figure 2: Example of HPAT Transfer Rules

references). This similarity is measured by N-gram precision scores. Several kinds of N-grams can be used in BLEU. We use from 1-gram to 4-gram in this paper, where a 1-gram precision score indicates the adequacy of word translation and longer N-gram (e.g., 4-gram) precision scores indicate fluency of sentence translation. The BLEU score is calculated from the product of N-gram precision scores, so this measure combines adequacy and fluency.

Note that a sizeable set of MT results is necessary in order to calculate an accurate BLEU score. Although it is possible to calculate the BLEU score of a single MT result, it contains errors from the subjective evaluation. BLEU cancels out individual errors by summing the similarities of MT results. Therefore, we need all of the MT results from the evaluation corpus in order to calculate an accurate BLEU score.

One feature of BLEU is its use of multiple references for a single source sentence. However, one reference per sentence is used in this paper because an already existing bilingual corpus is applied to the cleaning.

4 Feedback Cleaning

In this section, we introduce the proposed method, called feedback cleaning. This method is carried out by selecting or removing translation rules to increase the BLEU score of the evaluation corpus (Figure 1). Thus, this task is regarded as a combinatorial optimization problem of translation rules. The hill-climbing algorithm, which involves the features of this task, is applied to the optimization. The following sections describe the reasons for using this method and its procedure. The hill-climbing algorithm often falls into locally optimal solutions. However, we believe that a locally optimal solution is more effective in improving MT quality than the previous methods.

4.1 Costs of Combinatorial Optimization

Most combinatorial optimization methods iterate changes in the combination and the evaluation. In the machine translation task, the evaluation process requires the longest time. For example, in order to calculate the BLEU score of a combination (solution), we have to translate C times, where C denotes the size of the evaluation corpus. Furthermore, in order to find the nearest neighbor solution, we have to calculate all BLEU scores of the neighborhood. If the number of rules is R and the neighborhood is regarded as consisting of combinations made by changing only one rule, we have to translate $C \times R$ times to find the nearest neighbor solution. Assume that $C = 10,000$ and $R = 100,000$, the number of sentence translations (sentences to be translated) becomes one billion. It is infeasible to search for the optimal solution without reducing the number of sentence translations.

A feature of this task is that removing rules is easier than adding rules. The rules used for translating a sentence can be identified during the translation. Conversely, the source sentence set $S[r]$, where a rule r is used for the translation, is determined once the evaluation corpus is translated. When r is removed, only the MT results of $S[r]$ will change, so we do not need to re-translate other sentences. Assuming that five rules on average are applied to translate a sentence, the number of sentence translations becomes $5 \times C + C = 60,000$ for testing all rules. On the contrary, to add a rule, the entire corpus must be re-translated because it is unknown which MT results will change by adding a rule.

4.2 Cleaning Procedure

Based on the above discussion, we utilize the hill-climbing algorithm, in which the initial solution contains all rules (called the base rule set) and the search for a combination is done by only removing

static: C_{eval}, an evaluation corpus
R_{base}, a rule set acquired from the entire training corpus (the base rule set)
R, a current rule set, a subset of the base rule set
$S[r]$, a source sentence set where the rule r is used for the translation
Doc_{iter}, an MT result set of the evaluation corpus translated with the current rule set

---

procedure CLEAN-RULESET ()
  $R \leftarrow R_{base}$
  repeat
    $R_{iter} \leftarrow R$
    $R_{remove} \leftarrow \emptyset$
    $score_{iter} \leftarrow$ SET-TRANSLATION()
    for each r **in** R_{iter} **do**
      if $S[r] \neq \emptyset$ **then**
        $R \leftarrow R_{iter} - \{r\}$
        translate all sentences in $S[r]$, and obtain the MT results $T[r]$
        $Doc[r] \leftarrow$ the MT result set that $T[r]$ is replaced from Doc_{iter}
        the rule contribution $contrib[r] \leftarrow score_{iter} -$ BLEU-SCORE($Doc[r]$)
        if $contrib[r] < 0$ **then** add r to R_{remove}
    end
    $R \leftarrow R_{iter} - R_{remove}$
  until $R_{remove} = \emptyset$

---

function SET-TRANSLATION () **returns** a BLEU score of the evaluation corpus translated with R
  $Doc_{iter} \leftarrow \emptyset$
  for each r **in** R_{base} **do** $S[r] \leftarrow \emptyset$ **end**
  for each s **in** C_{eval} **do**
    translate s and obtain the MT result t
    obtain the rule set $R[s]$ that is used for translating s
    for each r **in** $R[s]$ **do** add s to $S[r]$ **end**
    add t to Doc_{iter}
  end
  return BLEU-SCORE(Doc_{iter})

Figure 3: Feedback Cleaning Algorithm

rules. The algorithm is shown in Figure 3. This algorithm can be summarized as follows.

- Translate the evaluation corpus first and then obtain the rules used for the translation and the BLEU score before removing rules.
- For each rule one-by-one, calculate the BLEU score after removing the rule and obtain the difference between this score and the score before the rule was removed. This difference is called the **rule contribution**.
- If the rule contribution is negative (i.e., the BLUE score increases after removing the rule), remove the rule.

In order to achieve faster convergence, this algorithm removes all rules whose rule contribution is negative in one iteration. This assumes that the removed rules are independent from one another.

5 N-fold Cross-cleaning

In general, most evaluation corpora are smaller than training corpora. Therefore, omissions of cleaning

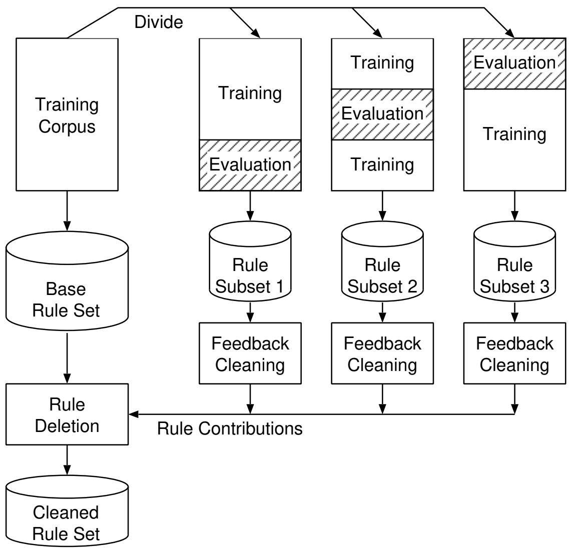

Figure 4: Structure of Cross-cleaning
(In the case of three-fold cross-cleaning)

will remain because not all rules can be tested by the evaluation corpus. In order to avoid this problem, we propose an advanced method called **cross-cleaning** (Figure 4), which is similar to cross-validation.

The procedure of cross-cleaning is as follows.

1. First, create the base rule set from the entire training corpus.

2. Next, divide the training corpus into N pieces uniformly.

3. Leave one piece for the evaluation, acquire rules from the rest $(N-1)$ of the pieces, and repeat them N times. Thus, we obtain N pairs of rule set and evaluation sub-corpus. Each rule set is a subset of the base rule set.

4. Apply the feedback cleaning algorithm to each of the N pairs and record the rule contributions even if the rules are removed. The purpose of this step is to obtain the rule contributions.

5. For each rule in the base rule set, sum up the rule contributions obtained from the rule subsets. If the sum is negative, remove the rule from the base rule set.

The major difference of this method from cross-validation is Step 5. In the case of cross-cleaning, the rule subsets cannot be directly merged because some rules have already been removed in Step 4. Therefore, we only obtain the rule contributions from the rule subsets and sum them up. The summed contribution is an approximate value of the rule contribution to the entire training corpus. Cross-cleaning removes the rules from the base rule set based on this approximate contribution.

Cross-cleaning uses all sentences in the training corpus, so it is nearly equivalent to applying a large evaluation corpus to feedback cleaning, even though it does not require specific evaluation corpora.

Set Name	Feature	English	Japanese
Training Corpus	# of Sentences	149,882	
	# of Words	868,087	984,197
Evaluation Corpus	# of Sentences	10,145	
	# of Words	59,533	67,554
Test Corpus	# of Sentences	10,150	
	# of Words	59,232	67,193

Table 1: Corpus Size

6 Evaluation

In this section, the effects of feedback cleaning are evaluated by using English-to-Japanese translation.

6.1 Experimental Settings

Bilingual Corpora The corpus used in the following experiments is the Basic Travel Expression Corpus (Takezawa et al., 2002). This is a collection of Japanese sentences and their English translations based on expressions that are usually found in phrasebooks for foreign tourists. We divided it into sub-corpora for training, evaluation, and test as shown in Table 1. The number of rules acquired from the training corpus (the base rule set size) was 105,588.

Evaluation Methods of MT Quality We used the following two methods to evaluate MT quality.

1. **Test Corpus BLEU Score**
   The BLUE score was calculated with the test corpus. The number of references was one for each sentence, in the same way used for the feedback cleaning.

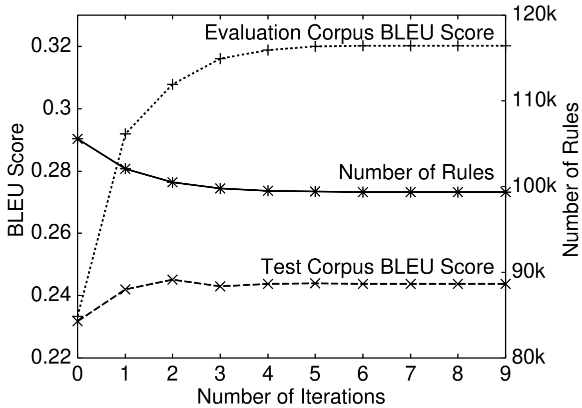

Figure 5: Relationship between Number of Iterations and BLEU Scores/Number of Rules

2. **Subjective Quality**
   A total of 510 sentences from the test corpus were evaluated by paired comparison. Specifically, the source sentences were translated using the base rule set, and the same sources were translated using the rules after the cleaning. One-by-one, a Japanese native speaker judged which MT result was better or that they were of the same quality. Subjective quality is represented by the following equation, where I denotes the number of improved sentences and D denotes the number of degraded sentences.

$$\text{Subj. Quality} = \frac{I - D}{\text{\# of test sentences}} \quad (1)$$

6.2 Feedback Cleaning Using Evaluation Corpus

In order to observe the characteristics of feedback cleaning, cleaning of the base rule set was carried out by using the evaluation corpus. The results are shown in Figure 5. This graph shows changes in the test corpus BLEU score, the evaluation corpus BLEU score, and the number of rules along with the number of iterations.

Consequently, the removed rules converged at nine iterations, and 6,220 rules were removed. The evaluation corpus BLEU score was improved by increasing the number of iterations, demonstrating that the combinatorial optimization by the hill-climbing algorithm worked effectively. The test corpus BLEU score reached a peak score of 0.245 at the second iteration and slightly decreased after the third iteration due to overfitting. However, the final score was 0.244, which is almost the same as the peak score.

The test corpus BLEU score was lower than the evaluation corpus BLEU score because the rules used in the test corpus were not exhaustively checked by the evaluation corpus. If the evaluation corpus size could be expanded, the test corpus score would improve.

About 37,000 sentences were translated on average in each iteration. This means that the time for an iteration is estimated at about ten hours if translation speed is one second per sentence. This is a short enough time for us because our method does not require real-time processing. [2]

6.3 MT Quality vs. Cleaning Methods

Next, in order to compare the proposed methods with the previous methods, the MT quality achieved by each of the following five methods was measured.

1. **Baseline**
   The MT results using the base rule set.
2. **Cutoff by Frequency**
   Low-frequency rules that appeared in the training corpus less often than twice were removed from the base rule set. This threshold was experimentally determined by the test corpus BLEU score.
3. χ^2 **Test**
   The χ^2 test was performed in the same manner as in Imamura (2002)'s experiment. We introduced rules with more than 95 percent confidence ($\chi^2 \geq 3.841$).
4. **Simple Feedback Cleaning**
   Feedback cleaning was carried out using the evaluation corpus in Table 1.
5. **Cross-cleaning**
   N-fold cross-cleaning was carried out. We applied five-fold cross-cleaning in this experiment.

The results are shown in Table 2. This table shows that the test corpus BLEU score and the subjective

[2] In this experiment, it took about 80 hours until convergence using a Pentium 4 2-GHz computer.

	Baseline	Previous Methods		Proposed Methods	
		Cutoff by Freq.	χ^2 Test	Simple FC	Cross-cleaning
# of Rules	105,588	26,053	1,499	99,368	82,462
Test Corpus BLEU Score	0.232	0.234	0.157	**0.244**	**0.277**
Subjective Quality		+1.77%	-6.67%	**+6.67%**	**+10.0%**
# of Improved Sentences		83	115	83	100
# of Same Quality		353	246	378	361
(Same Results)		(257)	(114)	(266)	(234)
# of Degraded Sentences		74	149	49	49

Table 2: MT Quality vs. Cleaning Methods

quality of the proposed methods (simple feedback cleaning and cross-cleaning) are considerably improved over those of the previous methods.

Focusing on the subjective quality of the proposed methods, some MT results were degraded from the baseline due to the removal of rules. However, the subjective quality levels were relatively improved because our methods aim to increase the portion of correct MT results.

Focusing on the number of the rules, the rule set of the simple feedback cleaning is clearly a locally optimal solution, since the number of rules is more than that of cross-cleaning, although the BLEU score is lower. In comparing the number of rules in cross-cleaning with that in the cutoff by frequency, the former is three times higher than the latter. We assume that the solution of cross-cleaning is also the locally optimal solution. If we could find the globally optimal solution, the MT quality would certainly improve further.

7 Discussion

7.1 Other Automatic Evaluation Methods

The idea of feedback cleaning is independent of BLEU. Some automatic evaluation methods of MT quality other than BLEU have been proposed. For example, Su et al. (1992), Yasuda et al. (2001), and Akiba et al. (2001) measure similarity between MT results and the references by DP matching (edit distances) and then output the evaluation scores. These automatic evaluation methods that output scores are applicable to feedback cleaning.

The characteristics common to these methods, including BLEU, is that the similarity to references are measured for each sentence, and the evaluation score of an MT system is calculated by aggregating the similarities. Therefore, MT results of the evaluation corpus are necessary to evaluate the system, and reducing the number of sentence translations is an important technique for all of these methods.

The effects of feedback cleaning depend on the characteristics of objective measures. DP-based measures and BLEU have different characteristics (Yasuda et al., 2003). The exploration of several measures for feedback cleaning remains an interesting future work.

7.2 Domain Adaptation

When applying corpus-based machine translation to a different domain, bilingual corpora of the new domain are necessary. However, the sizes of the new corpora are generally smaller than that of the original corpus because the collection of bilingual sentences requires a high cost.

The feedback cleaning proposed in this paper can be interpreted as adapting the translation rules so that the MT results become similar to the evaluation corpus. Therefore, if we regard the bilingual corpus of the new domain as the evaluation corpus and carry out feedback cleaning, the rule set will be adapted to the new domain. In other words, our method can be applied to adaptation of an MT system by using a smaller corpus of the new domain.

8 Conclusions

In this paper, we proposed a feedback cleaning method that utilizes automatic evaluation to remove incorrect/redundant translation rules. BLEU was utilized for the automatic evaluation of MT quality, and the hill-climbing algorithm was applied to searching for the combinatorial optimization. Utilizing features of this task, incorrect/redundant rules were removed from the initial solution, which contains all rules acquired from the training corpus. In addition, we proposed N-fold cross-cleaning to reduce the influence of the evaluation corpus size. Our experiments show that the MT quality was improved by 10% in paired comparison and by 0.045 in the BLEU score. This is considerable improvement over the previous methods.

Acknowledgment

The research reported here is supported in part by a contract with the Telecommunications Advancement Organization of Japan entitled, "A study of speech dialogue translation technology based on a large corpus."

References

Yasuhiro Akiba, Kenji Imamura, and Eiichiro Sumita. 2001. Using multiple edit distances to automatically rank machine translation output. In *Proceedings of Machine Translation Summit VIII*, pages 15–20.

Osamu Furuse and Hitoshi Iida. 1994. Constituent boundary parsing for example-based machine translation. In *Proceedings of COLING-94*, pages 105–111.

Kenji Imamura. 2001. Hierarchical phrase alignment harmonized with parsing. In *Proceedings of the 6th Natural Language Processing Pacific Rim Symposium (NLPRS 2001)*, pages 377–384.

Kenji Imamura. 2002. Application of translation knowledge acquired by hierarchical phrase alignment for pattern-based MT. In *Proceedings of the 9th Conference on Theoretical and Methodological Issues in Machine Translation (TMI-2002)*, pages 74–84.

Arul Menezes and Stephen D. Richardson. 2001. A best first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora. In *Proceedings of the 'Workshop on Example-Based Machine Translation' in MT Summit VIII*, pages 35–42.

Adam Meyers, Michiko Kosaka, and Ralph Grishman. 2000. Chart-based translation rule application in machine translation. In *Proceedings of COLING-2000*, pages 537–543.

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)*, pages 311–318.

Keh-Yih Su, Ming-Wen Wu, and Jing-Shin Chang. 1992. A new quantitative quality measure for machine translation systems. In *Proceedings of COLING-92*, pages 433–439.

Toshiyuki Takezawa, Eiichiro Sumita, Fumiaki Sugaya, Hirofumi Yamamoto, and Seiichi Yamamoto. 2002. Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, pages 147–152.

Keiji Yasuda, Fumiaki Sugaya, Toshiyuki Takezawa, Seiichi Yamamoto, and Masuzo Yanagida. 2001. An automatic evaluation method of translation quality using translation answer candidates queried from a parallel corpus. In *Proceedings of Machine Translation Summit VIII*, pages 373–378.

Keiji Yasuda, Fumiaki Sugaya, Toshiyuki Takezawa, Seiichi Yamamoto, and Masuzo Yanagida. 2003. Applications of automatic evaluation methods to measuring a capability of speech translation system. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003)*, pages 371–378.