Regular Paper

Multiple Translation-Engine-based Hypotheses and Edit-Distance-based Rescoring for a Greedy Decoder for Statistical Machine Translation

Michael Paul,[†,††] Eiichiro Sumita[†] and Seiichi Yamamoto[†,†††]

This paper extends a greedy decoder for statistical machine translation (SMT), which searches for an optimal translation by using SMT models starting from a decoder seed, i.e., the source language input paired with an initial translation hypothesis. First, the outputs generated by *multiple translation engines* are utilized as the initial translation hypotheses, whereby their variations reduce *local optima* problems inherent in the search. Second, a *rescoring method* based on the edit-distance between the initial translation hypothesis and the outputs of the decoder is used to compensate for problems of conventional greedy decoding solely based on statistical models. Our approach is evaluated for the translation of dialogues in the travel domain, and the results show that it drastically improves translation quality.

1. Introduction

Statistical approaches to machine translation (MT) have achieved much progress over the last decade. This paper focuses on one of the state-of-the-art approaches, i.e., the *greedy decoding* approach described in Section 2. Despite a high performance on average, the greedy decoding approach can often produce translations with severe errors.

This paper addresses two problems of the greedy decoding approach:

(1) The greedy decoder searches for the translation that is most likely starting from a decoder seed, i.e., the source language input paired with an initial translation hypothesis. The selection of the starting point is crucial to avoid local optima in the search. However, this problem has not yet been addressed much.
(2) The greedy decoder generates multiple translations out of which a single translation is selected according to its statistical models. However, the selected one is not necessarily the best-quality translation.

To solve these two problems, Section 3 extends the greedy decoding approach as follows:

First, we focus on the starting point problem. We propose a method of using diverse starting points generated by multiple translation engines. Combining multiple MT systems has the advantage of exploiting the strengths of each MT engine. Quite different initial translation hypotheses are produced due to particular output characteristics of each MT engine. Therefore, larger parts of the search space can be explored while avoiding local optima problems of the search algorithm.

Second, we propose an *edit-distance-based rescoring method* that addresses the translation selection problem of conventional greedy decoding solely based on statistical models. The rescoring algorithm compares the initial translation hypothesis and the generated translations by using an *edit-distance* measure. The edit-distance is combined with the statistical scores to select the best-quality translation.

The effects of the proposed method are demonstrated in Section 4 for the Japanese-to-English translation of dialogues in the travel domain.

2. Greedy Decoding for SMT

In this section, we explain the outline of SMT and greedy decoding in short.

2.1 Statistical Machine Translation

Statistical machine translation formulates the problem of translating a sentence from a source language S into a target language T as the maximization problem:

$$\text{argmax}_T \; p(S|T) * p(T), \qquad (1)$$

where $p(S|T)$ is called a *translation model* (TM), representing the generation probability from T into S, and $p(T)$ is called a *language model* (LM), which represents the likelihood of the target language [2]. During the translation

† ATR Spoken Language Communication Research Laboratories
†† Graduate School of Science and Technology, Kobe University
††† Department of Information Systems Design, Doshisha University

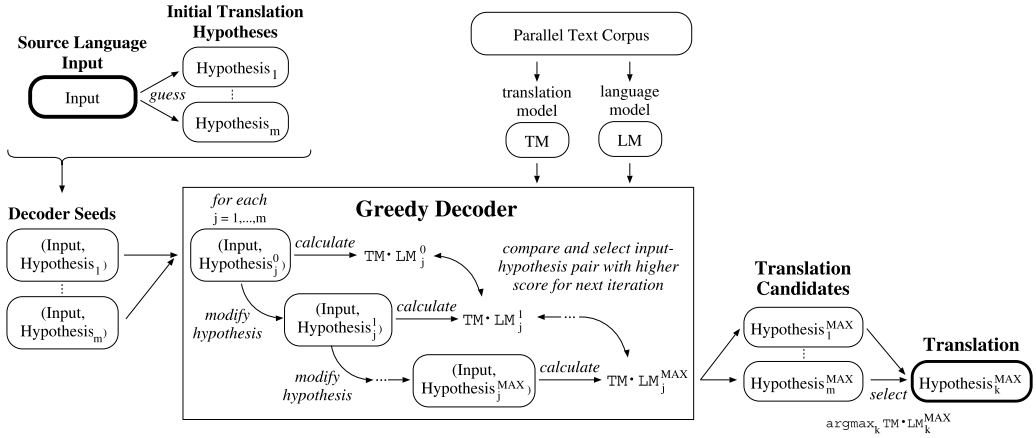

Fig. 2 Greedy decoding.

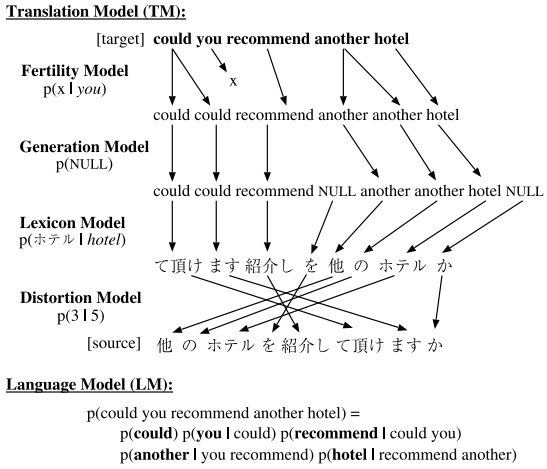

Fig. 1 Statistical models.

process (*decoding*), a statistical score based on TM and LM is assigned to each translation. In this paper, we call this score **TM·LM**. The translation with the highest TM·LM score is selected as the output.

We used the *IBM-4* translation model [2] in the experiments in Section 4, which consists of probabilities for word translations (*lexicon model*), the number of source words produced by a target word (*fertility model*), word insertions (*generation model*), and word order changes (*distortion model*). LM is based on the frequency of consecutive word sequences (*n-gram*). The TM and LM probabilities are trained automatically from a parallel text corpus.

Figure 1 gives an example for the process of transferring a Japanese source sentence into an English target sentence and illustrates which translation knowledge is captured by the respective statistical models mentioned above.

2.2 Greedy Decoding

Various decoding algorithms have been proposed, including *stack-based* [21], *beam search* [18], and *greedy decoding* [5]. This paper concentrates on the greedy decoding approach described in details in Section 2.2.1. Problems of this approach are summarized in Section 2.2.2.

2.2.1 Algorithm

Figure 2 illustrates the decoding algorithm, which is described in detail in German, et al.[5], and summarizes the terminology used throughout this paper.

The input of the decoder (*decoder seed*) consists of the input, i.e., the source language sentence, paired with an initial translation hypothesis, whereby the initial translation hypothesis is formed by a word-by-word translation of the source language sentence. The following steps attempt to improve the quality of the translation hypothesis by greedily exploring alternative translations starting from the initial translation hypothesis. The algorithm modifies the hypothesis iteratively using a set of word operations [5] such as *inserting*, *deleting*, *joining*, and *swapping*. After each modification, the statistical scores of the previous and modified input-hypothesis pairs are calculated. If the modified input-hypothesis pair has a higher TM·LM score, it is used in the next iteration. Otherwise, the modified hypothesis is ignored and the search is continued using the previous input-hypothesis pair. The decoding algorithm stops if no further improvement can be achieved by any operation and outputs the hypothesis with the *highest statistical score*.

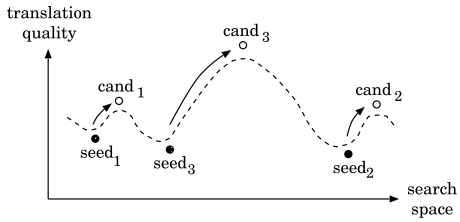

Fig. 3 Local optima problem of the greedy search.

If multiple initial translation hypotheses are used for a given source language input, the decoder is applied to each of the initial translation hypotheses, resulting in multiple translation candidates, and the candidate with the highest statistical score is selected as the translation.

2.2.2 Two Major Problems of Greedy Decoding

The greedy decoding approach has two major problems:

(1) The translation output depends on the initial translation hypothesis to start the search, which may lead to a local optimum translation but not to the global optimum translation.
(2) The best-quality translation in the list of translation candidates isn't selected as the final output when lower statistical TM·LM scores are assigned, despite of good quality.

Figure 3 illustrates the first problem. Given the decoder seed $seed_1$, the greedy decoder modifies the initial translation hypothesis based on its statistical models (along the dotted line) as long as the TM·LM score increases and finally outputs the translation candidate with maximal score ($cand_1$). Similarly, the local optimum translation candidate $cand_2$ is obtained when $seed_2$ is used as the decoder seed. However, using $seed_3$ as the starting point, the decoder finds the global optimum translation candidate $cand_3$ that cannot be found by using the other seeds. Word-by-word translation (cf. Section 2.2.1) often fails to produce decoder seeds like $seed_3$.

The second problem of the greedy decoding approach is the selection of worse translation candidates according to its statistical models. This occurs partly because the decoder might modify hypotheses wrongly resulting in translations of lower quality with higher statistical scores.

Both problems are demonstrated with experiments in Section 4.2.

3. Multi-Engine-based Hypotheses and Edit-Distance-based Rescoring

We propose solving the two problems of greedy decoding by:

(1) using *multiple translation-engine-based decoder seeds* to start the search
(2) using an *edit-distance-based rescoring method* to select the best translation

First, our approach utilizes translations produced by multiple translation engines as the initial translation hypotheses. The multi-engine approach has the advantage of exploiting the strengths of each MT engine. Due to the particular output characteristics of each MT engine, quite different initial translation hypotheses are produced. Therefore, larger parts of the search space can be explored while avoiding the local optima problem in the search (cf. Section 3.1).

Second, we propose an edit-distance-based rescoring method. The rescoring algorithm checks the difference between the initial translation hypothesis and the generated translation candidates using *edit-distance*. The edit-distance is combined with the statistical scores in order to compensate for problems of conventional greedy decoding (cf. Section 3.2).

The integrated proposal is outlined in Section 3.3.

3.1 Multi-Engine-based Hypotheses

Various methods can be utilized to produce initial translation hypotheses. In the case of the original greedy decoding, the initial translation hypothesis is obtained as a word-by-word translation of the source language input.

In place of such a dictionary-based approach, example-based methods can be exploited. For example, Watanabe and Sumita [22] proposed retrieving translation examples, i.e., pairs of a source and a human-translated target language sentence whose source sentences are similar to the input sentence, from a parallel text corpus and then using the respective target language sentences as initial translation hypotheses.

In this paper, we propose a method of translating the source language input by available translation engines and using the obtained MT outputs as the initial translation hypotheses. For the experiments and discussions given in this paper, we focus on the following two methods to generate initial translation hypotheses:

(1) *Previous Method*:
an *example-based* method (*EB*) that ex-

tracts an initial translation hypothesis from a parallel text corpus (cf. Section 3.1.1).

(2) *Proposed Method*:
an *MT-based* method that uses an MT engine to translate the input and takes the MT output as the initial translation hypothesis (cf. Section 3.1.2).

The examples given below are taken from the Japanese-to-English translation experiments described in Section 4.

3.1.1 Example-based Hypotheses

Watanabe and Sumita[22] proposed an example-based method that utilizes the *tf·idf* criteria[10] as seen in the information retrieval framework to extract the most similar translation examples for a given source language input from a parallel text corpus. The target language sentence of the translation example with the highest similarity score is selected as the initial translation hypothesis. If multiple translation examples obtain the same highest score, all target parts are used as initial translation hypotheses.

Depending on the coverage of the training corpus, the number of hypotheses retrieved for a given input sentence might vary. A large number of hypotheses is retrieved by the example-based method, if the input is short. The longer the input, the fewer hypotheses can be extracted. However, at least one hypothesis, i.e., the one with the highest similarity score, is retrieved. In the experiments described in Section 4, the number of retrieved hypotheses for a given input varied between 1 and 119, and single hypotheses were obtained for 62.7% of the input sentences. On average, six hypotheses were retrieved. Samples of the example-based hypotheses are given in **Table 1**.

3.1.2 MT-based Hypotheses

The MT-based method proposed in this paper utilizes the output of a translation engine as the initial translation hypothesis. For our experiments, we used the seven MT engines listed in **Table 2**☆.

Two of them (MT_{1-2}) are in-house *example-based MT* (EBMT) systems that are trained on the same training set as the greedy decoder. The remaining five (MT_{3-7}) are *off-the-shelf MT* (OTSMT) systems that are based on lexicons, grammars, and translation rules. Examples of MT-based hypotheses for the input sentences of Table 1 are given in **Table 3**.

☆ The MT engines are listed alphabetically, where the order is unrelated to the indexing scheme (MT_i) used for the examples and the discussion of the evaluation results given in this paper.

Table 1 Example-based hypotheses.

(source language input)
しょうゆをお願いできますか
(→ *do you have any soy sauce*)
(initial translation hypothesis)
can i ask for a guide
can i exchange money
can i have room service please
can i have tea with lemon
can i have the check
can you order it for me
could i speak to your sales manager
do you have japanese tea
excuse me can i order now
may i have a blanket
may i have a wake up call
(source language input)
ハイアットリージェンシーホテルをお願いします シングルルームに泊まりたいのですが
(→ *i would prefer the hyatt regency please and if possible i want a single room*)
(initial translation hypothesis)
could you reserve a single room for me

Table 2 Utilized MT engines.

EBMT	(in-house)	D3[15]
	(in-house)	HPAT[7]
OTSMT	Fujitsu	ATLAS[4]
	NEC	CROSSROAD[12]
	IBM	HONYAKUOOSAMA[6]
	LogoVista	LOGOVISTA[9]
	Toshiba	THEHONYAKU[19]

3.1.3 Characteristics of Example-based and MT-based Initial Translation Hypotheses

The initial translation hypotheses obtained by the example-based and the MT-based methods have quite different characteristics. We analyzed the initial translation hypotheses obtained for the experiments described in Section 4 to clarify how much they differ from each other by using the edit-distance measure defined in Section 3.2.1.

For each of the N input sentences of the test set, we examined the set S_i ($1 \le i \le N$) consisting of the m_i initial translation hypotheses used in the greedy decoding of the input sentences. For each set S_i, we calculated the average edit-distance between all pairs $\{H_j, H_k\}$ of initial translation hypotheses contained in the same set (H_j, $H_k \in S_i$; $1 \le j < k \le m_i$). In the case of a single initial translation hypothesis ($m_i = 1$), a score of zero was used instead.

In order to compare the initial translation hypotheses of the example-based and the MT-based methods, we calculated the average edit-

Table 3 MT-based hypotheses.

(source language input)
しょうゆをお願いできますか
(→ *do you have any soy sauce*)
(initial translation hypothesis)
MT_1: do you have soy sauce
MT_2: can i have the soy sauce please
MT_3: please give me soy sauce
MT_4: could the soy sauce be done
MT_5: can you ask for soy sauce
MT_6: please give me soy sauce
MT_7: is it possible to request soy sauce
(source language input)
ハイアットリージェンシーホテルをお願いします シングルルームに泊まりたいのですが
(→ *i would prefer the hyatt regency please and if possible i want a single room*)
(initial translation hypothesis)
MT_1: i 'm afraid i don't
MT_2: at the hyatt regency hotel please i 'd like to stay in a single room
MT_3: i want to stay at the single room which asks you for the hyatt regency hotel
MT_4: i want to stay at a single room in which it asks for the hyatt regency hotel
MT_5: i want to stay at the single room which you may ask for hyatt regency hotel with
MT_6: although he wants to stay at the single room which asks you for the hyatt regency hotel
MT_7: but wanting to stay at the single room to request hyatt regency hotel of

distance (ED_{avg}) over all hypothesis sets of each method as follows:

$$ED_{avg} = \frac{\left(\sum_{i=1}^{N} ED^{i}_{avg}\right)}{N},$$

$$ED^{i}_{avg} = \begin{cases} \dfrac{\sum_{j=1}^{m_i}\sum_{k=j+1}^{m_i} ED(H_j, H_k)}{\frac{1}{2}\cdot(m_i-1)\cdot m_i}, & if\ m_i > 1 \\ 0, & otherwise \end{cases}.$$

We obtained average scores of 1.5☆ for the example-based method and 8.7 for the combination of all MT-based hypotheses. These results show that the example-based method retrieves either a single hypothesis or hypotheses that are quite similar to each other, because there might be only a few variations in the expressions covered by the training corpus. On the other hand, the MT-based hypotheses show larger variations, because they are produced by independently developed translation engines that use different dictionaries, grammars, and translation rules.

This indicates that the decoding of multiple example-based hypotheses might result in similar decoder outputs, while the decoding of the MT-based hypotheses may provide translations that result in various outputs, increasing the chance to catch the global optimum (cf. Figure 3).

☆ As mentioned in Section 3.1.1, the example-based method retrieved 62.7% of single hypotheses. The average score for the remaining sentences with multiple hypotheses is 3.9.

3.2 Edit-Distance-based Rescoring

In order to address the second problem of conventional greedy decoders, i.e., the selection of bad translation candidates with high statistical scores, we propose an *edit-distance-based rescoring method* that compensates the statistical scores of each generated translation candidate by information on how much the initial translation hypothesis is modified during decoding based on the costs of edit-operations (cf. Section 3.2.1). *The more modifications* that are necessary to alter the initial translation hypothesis to the translation candidate, *the more likely it is that the candidate is a translation of bad quality.* The rescoring function is defined in Section 3.2.2.

3.2.1 Edit-Distance

The *edit-distance*[20] is a popular approach to measuring the distance between sequences of words. The distance is defined as the sum of the costs of *insertion* (INS), *deletion* (DEL), and *substitution* (SUB) operations required to map one word sequence (s_1) into the other (s_2).

$$\mathrm{ED}(s_1, s_2) = |\mathrm{INS}| + |\mathrm{DEL}| + |\mathrm{SUB}|,$$

where "$|x|$" states how many times operation x is applied. The edit-distance can be calculated by a *dynamic programming* technique[3].

3.2.2 Rescoring Function

The rescoring function *rescore* takes into account the TM·LM score of the translation candidate C that is generated by the decoder from the source language input I and the edit-distance between the translation candidate C and the initial translation hypothesis H.

$$rescore(C,I,H) = \mathrm{func}(\mathrm{TM{\cdot}LM}(C, I), \mathrm{ED}(C, H)).$$

If the initial translation hypothesis is already close to a correct translation, not many operations should be required. Therefore, not only *large* TM·LM(C, I) scores, but also *small* ED(C, H) scores are indicators of high quality translations.

The *rescore* function has to be designed in such a way that less-altered translation candidates with high translation and language model scores are preferred.

For the experiments described in this paper, we used the edit-distance of the translation candidate C and the initial translation hypothesis H as a weight to decrease the statistical scores, whereby the scaling factor *scale* is optimized on

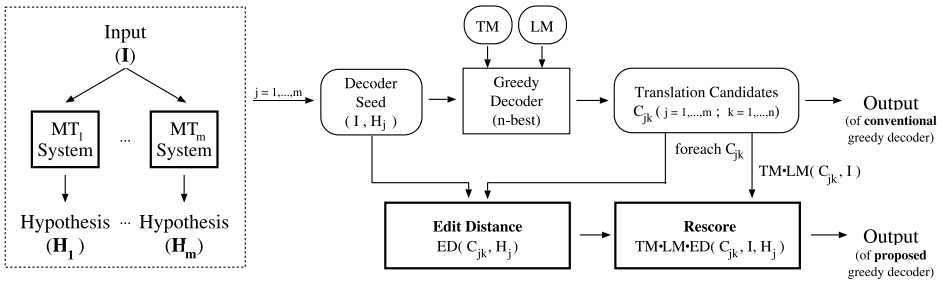

Fig. 4 Greedy decoding using multiple translation-engine-based hypotheses and edit-distance-based rescoring.

```
(1)  proc greedy-decode-MT-output-with-rescoring( Input, Corpus, MT_1, ..., MT_m ) ;
(2)  begin
(3)    TM ← translation-model(Corpus) ;                    (* initialize statistical models *)
(4)    LM ← language-model(Corpus) ;
(5)    HypList ← {} ;                                      (* create initial translation hypotheses *)
(6)    for each MT in {MT_1, ..., MT_m} do
(7)      HypList ← HypList ∪ translate(Input, MT) ;
(8)    od ;
(9)    CandList ← {} ;                                     (* apply greedy decoder *)
(10)   for each Hyp in HypList do
(11)     NbestList ← greedy-decode({Hyp, Input}, TM, LM) ;
(12)     for each Cand in NbestList do
(13)       CandList ← CandList ∪ {Cand, Hyp} ;
(14)     od ;
(15)   od ;
(16)   rescored-CandList ← {} ;                            (* apply rescoring method *)
(17)   for each {Cand, Hyp} in CandList do
(18)     rescored-CandList ← {Cand, TM·LM·ED(Cand, Input, Hyp)} ;
(19)   od ;
(20)   return( top(rescored-CandList) ) ;
(21) end ;
```

Fig. 5 Proposed algorithm.

a held-out set as described in Section 4.3.2.

$$\textbf{TM·LM·ED}(\mathrm{C,I,H}) = \frac{\mathrm{TM{\cdot}LM}(C,I)}{\exp(\ scale * \mathrm{ED}(C,H)\ )} . \quad (2)$$

If the initial translation hypothesis H cannot be improved by the greedy decoder according to its statistical models, the edit-distance is zero (C=H) and the revised score is identical to the TM·LM score.

3.3 Integration of Multiple Translation-Engine-based Hypotheses and Edit-Distance-based Rescoring

Figure 4 illustrates the flow of information in the proposed framework. Given a source language input I, multiple MT systems (MT_1,...,MT_m) are used to produce m initial translation hypotheses. Each of the initial translation hypotheses is paired with the source language input and used as the decoder seed of the greedy decoder. The decoder output consists of m n-best [☆] lists of translation candidates that are concatenated and ranked according to the statistical TM·LM score.

In the case of the conventional greedy decoding method, the translation candidate with the highest TM·LM score is selected as the translation output.

Within the proposed framework, all translation candidates (C_{jk}; $1 \le j \le \mathrm{m}$; $1 \le k \le \mathrm{n}$) are rescored using the TM·LM·ED function proposed in Section 3.2.2, and the translation candidate with the highest TM·LM·ED score is selected as the translation output. The proposed algorithm is summarized in **Fig. 5**.

4. Evaluation

The outline of the evaluation is summarized in **Fig. 6**. Section 4.1 describes the experimental setting. In order to train the translation [☆☆]

[☆] For the experiments described in Section 4, we used m=7 and n=10.

[☆☆] The translation models are trainedusing the GIZA++ toolkit, http://www.fjoch.com

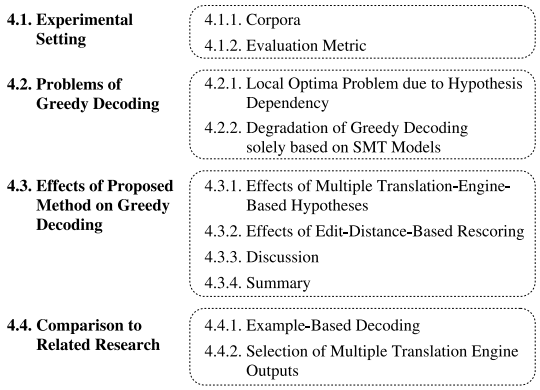

Fig. 6 Evaluation outline.

and language ☆ models, we use two corpora from the *travel* domain (cf. Section 4.1.1). The proposed method is evaluated by using a human assessment of *translation accuracy* which is defined in Section 4.1.2.

The problems of conventional greedy approaches are demonstrated experimentally in Section 4.2. First, the local optima problem due to hypothesis dependency is shown in Section 4.2.1. Second, the degradation problem of conventional greedy decoding approaches is illustrated in Section 4.2.2.

The effectiveness of the proposed method is investigated in Section 4.3. Section 4.3.1 shows the effects of using multiple MT-based initial translation hypotheses and Section 4.3.2 shows the effects of the edit-distance-based rescoring. The obtained results are discussed in Section 4.3.3 and summarized in Section 4.3.4.

Finally, Section 4.4 compares the performance of the proposed methods with related research, i.e., a previous greedy decoder [22] that relies solely on examples and statistical model scores (cf. Section 4.4.1) and a previous method [1] that uses multiple language and translation model pairs to select the best translation among multiple MT outputs (cf. Section 4.4.2).

4.1 Experimental Setting

In this section, we describe the corpora and evaluation metric.

4.1.1 Corpora

The evaluation of our approach is carried out using two Japanese-English parallel corpora of the *travel* domain.

☆ The language models are trained using the CMU-Cambridge Statistical Language Modeling Toolkit v2, http://mi.eng.cam.ac.uk/~prc14/toolkit.html

Table 4 Corpus statistics.

corpus	sentence count	language	word tokens	word types	words per sentence
BTEC	162,318	Japanese	1,114,186	18,781	6.9
		English	952,300	12,404	5.9
MAD	4,894	Japanese	62,529	2,607	10.0
		English	57,500	2,158	10.3

BTEC – *Basic Travel Expression Corpus* [17]
The BTEC corpus is a large collection of sentences ☆☆ that bilingual travel experts consider useful for people going to or coming from countries with different languages. The BTEC sentences are not transcriptions of actual interactions, but were written by experts.

MAD – *Machine Aided Dialogue Corpus* [8]
The MAD corpus is a collection of dialogues between a native speaker of Japanese and a native speaker of English that is mediated by a speech-to-speech translation system.

The statistics of the corpora are given in **Table 4**, where *word token* refers to the number of words in the corpus and *word type* refers to the vocabulary size. Since the MAD corpus consists of dialogues, it contains more complex and compound sentences as well as filled pauses, resulting in longer sentences that are more difficult to translate.

The corpora were split randomly into three parts for the acquisition of translation knowledge (*training set*), parameter tuning (*development set*), and evaluation purposes (*test set*). For the experiments described below, we selected randomly 505 held-out sentences from the MAD corpus as the development set and 502 sentences from the MAD corpus as the test set. The remaining sentences of MAD and BTEC were used for the training of the statistical models and the retrieval of initial translation hypotheses by the example-based method.

In other words, MAD is the target of our speech-to-speech translation system and BTEC is used as a resource to acquire translation knowledge for the translation system.

4.1.2 Evaluation Metric

For the evaluation of the *translation accuracy* [16] we use a human assessment. For each translation, a native speaker of the target language assigns ranks ranging from A to D (A:

☆☆ Parts of the BTEC corpus were used in the International Workshop of Spoken Language Translation (http://www.slt.atr.jp/IWSLT2004/) and will be made publicly available through GSK (http://www.gsk.or.jp/).

perfect translation, B: *fair* translation, C: *acceptable* translation, D: *nonsense*). Hereafter, we use the percentage of translations ranked A, B, or C as the *ABC* score, where higher ABC scores indicate better translations.

4.2 Problems of Greedy Decoding

In this section, we demonstrate experimentally the problems of conventional greedy approaches. Section 4.2.1 illustrates the local optima problem due to the dependency on the initial translation hypothesis. Section 4.2.2 demonstrates the degradation problem of greedy decoding relying solely on statistical models.

4.2.1 Local Optima Problem due to Hypothesis Dependency

In order to investigate the local optima problem due to the dependency of the greedy decoding approach on the initial translation hypothesis, we evaluated and compared the translation quality of the greedy decoder outputs when applied to the following types of initial translation hypotheses (cf. Section 3.1):

- all example-based hypotheses retrieved from the training corpus for a given input. We refer to this hypothesis type as EB_{all}.
- all MT-based hypotheses produced by the EBMT and OTSMT translation engines listed in Table 2. We refer to this hypothesis type as MT_{1-7}.

Table 5 summarizes the translation accuracy of the translations generated by the greedy decoder when applied to the respective hypothesis types, where the translation candidates are selected solely based on statistical models. The results show that:

- The translation quality of the greedy decoder differs largely for different types of initial translation hypotheses.
- The example-based retrieval method EB_{all} is outperformed by the MT-based method MT_{1-7} by 10% in translation accuracy.

As shown in Section 3.1.3, the initial translation hypotheses obtained by the example-based and MT-based methods have quite different characteristics, which affect the translation quality of the greedy decoder outputs. The effects of using initial translation hypotheses with larger variations are illustrated in **Table 6**. It summarizes the outputs of a greedy decoder solely based on statistical models when the example-based (cf. Table 1) and MT-based (cf. Table 3) initial translation hypotheses are used to translate the source language input "ハイアットリージェンシーホテルをお願いしますシングルルームに泊まりたいのですが".

Table 5 Local optima problem due to hypothesis dependency.

initial translation hypotheses		translation accuracy ABC (%)
EB	EB_{all}	53.9
EBMT+OTSMT	MT_{1-7}	63.9

Table 6 Greedy decoding outputs using different hypothesis types.

(source language input)
ハイアットリージェンシーホテルをお願いします
シングルルームに泊まりたいのですが
(→ *i would prefer the hyatt regency please and if possible i want a single room*)

example-based hypotheses
(cf. Table 1)

(translation)

EB_{all}:	[C]	i want a single room for me

MT-based hypotheses
(cf. Table 3)

(translation)

MT_1:	[D]	i 'm afraid
MT_2:	[A]	i 'd like to stay in a single room at the hyatt regency hotel please
MT_3:	[B]	i want to stay at the hyatt regency hotel single room
MT_4:	[C]	i want to stay in a single room
MT_5:	[B]	i want to stay at the hyatt regency hotel single room
MT_6:	[B]	i want to stay at the hyatt regency hotel single room
MT_7:	[B]	i want to stay at the hyatt regency hotel single room

"[X]" indicates the human assessment (cf. Section 4.1.2)

The example-based method retrieves a single hypothesis from the training corpus, which covers only parts of the information contained in the source language input. Therefore, no good translation can be generated.

The MT-based hypotheses, however, show larger variations, which increases the chance of overcoming the local optima problems in the search. Although four initial translation hypotheses ($\mathrm{MT}_{3,5,6,7}$) in the given example are decoded into the same local optimum translation candidate, a better translation is found by the MT-based method using the initial translation hypothesis of MT_2.

4.2.2 Degradation of Greedy Decoding solely based on SMT Models

In order to get an idea of how much degradation can be expected even with high-quality hypotheses, we used the correct human translations of the MAD test set as the initial translation hypotheses. The results show a large degradation, i.e., *30.3% of the human translations resulted in unacceptable translations*. Examples of degraded decoder outputs are given in **Table 7**.

Table 7 Degradation of conventional greedy decoding for perfect initial translation hypotheses.

(source language input)
私道に迷ったようなんですが道を教えていただけませんか
(initial translation hypothesis)
i think i am lost could you help me
(translation)
[B] i 'm lost could you tell me
(source language input)
ツインのお部屋を 1 人でお使いになられるのでしたら取れますが
(initial translation hypothesis)
if you 'll be staying alone i can reserve a twin room
(translation)
[C] i have a twin room alone if you be using
(source language input)
わかりましたありがとうございましたお手数お掛けしました
(initial translation hypothesis)
okay thanks a lot i apologize for the inconvenience
(translation)
[D] thank you for the inconvenience

"[X]" indicates the human assessment (cf. Section 4.1.2)

An analysis of the translation candidate lists generated by the decoder, however, revealed that most of the original human translations are left in the list, but with statistical scores assigned that are lower than those of the selected decoder output. These results indicate that *not only a good selection of the initial translation hypothesis but also a careful verification of the translation quality of the decoder output is required* to improve the performance of conventional greedy decoding approaches.

4.3 Effects of Proposed Method on Greedy Decoding

In this section, we investigate how the proposed method affects the performance of the greedy decoder. Section 4.3.1 shows the effects of multiple MT-based decoder seeds and Section 4.3.2 shows the effects of the edit-distance-based rescoring method. The obtained results are discussed in Section 4.3.3. Finally, the effects of combining the usage of multiple translation-engine-based hypotheses and edit-distance-based rescoring are summarized in Section 4.3.4.

4.3.1 Effects of Multiple Translation-Engine-based Hypotheses

In order to investigate the effects of using multiple MT-based decoder seeds for a given input, we compared the translation quality of single-seed decoder outputs with decoding results by using multiple seeds, where the translation output is selected based on statistical scores only, i.e., no rescoring is applied. The single-seed decoder outputs were generated from the respective MT-based hypotheses (MT_1, ..., MT_7). For the multi-seed decoder outputs, we evaluated the system performances with the EBMT-based hypotheses (MT_{1-2}), the OTSMT-based hypotheses (MT_{3-7}), and the combination of all MT-based hypotheses (MT_{1-7}).

Table 8 Effects of multiple translation-engine-based hypotheses.

Single MT-based seed

initial translation hypotheses		translation accuracy ABC (%)
EBMT	MT_1	37.2
	MT_2	63.7
OTSMT	MT_3	51.3
	MT_4	50.5
	MT_5	51.3
	MT_6	50.7
	MT_7	45.6

Multiple MT-based seed

initial translation hypotheses		translation accuracy ABC (%)
EBMT	MT_{1-2}	63.5
OTSMT	MT_{3-7}	58.1
EBMT +OTSMT	MT_{1-7}	63.9

Multiple Example-based seed

EB	EB_{all}	53.9

In addition, we compared the system performance using the proposed MT-based method (MT_{1-7}) with the system performance of using the example-based method (EB_{all}).

The comparison of the evaluation results summarized in **Table 8** shows that:

- The MT-based multi-seed systems achieved better results than the best single-seed systems whose initial translation hypotheses are used (MT_2 is an exception), thus showing *the effectiveness of using multiple MT-based seeds* as the input of the greedy decoder.
- The combination of EBMT-based and OTSMT-based initial translation hypotheses (MT_{1-7}) further improves the system performance.
- All MT-based multi-seed systems (MT_{1-2}, MT_{3-7}, MT_{1-7}) outperform the example-based seed system, thus showing *the potential of the MT-based initial translation hypotheses to avoid local optima problems in the search.*

4.3.2 Effects of Edit-Distance-based Rescoring

In order to investigate how the proposed rescoring function affects the performance of

Table 9 Effects of edit-distance-based rescoring.

Greedy decoding without rescoring (TM·LM)

initial translation hypotheses		translation accuracy ABC (%)
EBMT	MT_{1-2}	63.5
OTSMT	MT_{3-7}	58.1
EBMT +OTSMT	MT_{1-7}	63.9

Greedy decoding witht rescoring (TM·LM·ED)

initial translation hypotheses		translation accuracy ABC (%)
EBMT	MT^r_{1-2}	66.7
OTSMT	MT^r_{3-7}	65.1
EBMT +OTSMT	$\mathbf{MT^r_{1-7}}$	**73.1**

the greedy decoder, we first explain how the scaling factor of the rescoring function is determined. Next, we compare the translation quality of the proposed method using multiple seeds for a given input with the performance of a conventional greedy decoder solely based on statistical models.

4.3.2.a Scaling Factor Determination

The rescoring function TM·LM·ED defined in Section 3.2.2 includes a scaling factor that influences how much weight is given to the statistical score compared to the edit-distance score during the selection process.

We used a simple iterative method to determine the optimal scaling factor. The development set of MAD was translated by our method using the combination of all MT-based hypotheses (MT_{1-7}), with variable scaling factors ranging from 0 to 25. The obtained results were evaluated according to an automatic evaluation metric, i.e., the *word error rate*[14] (WER), which penalizes edit-operations for the translation output against reference translations[☆]. In contrast to ABC, smaller WER scores indicate better translations. The scaling factor that achieved the lowest WER score (*scale*=5.5) was used for the evaluation of the test set.

4.3.2.b Rescoring Function Contribution

Table 9 summarizes the effects of the TM·LM·ED rescoring function when applied to the list of translation candidates generated by the MT-based multi-seed systems introduced in Section 4.3.1, where "^{r}" indicates the usage of the rescoring function. A comparison of the results shows that:

- Greedy decoding with rescoring applied to multiple MT-based hypotheses outperforms conventional methods solely based on statistical models, thus showing *the potential to overcome the problem in translation candidate selection of conventional greedy decoding approaches.*
- A larger gain in performance is achieved for the rescoring of multiple OTSMT-based hypotheses (ABC: +7%) compared to multiple EBMT-based hypotheses (ABC: +3.2%).
- The combination of EBMT-based and OTSMT-based hypotheses (MT^r_{1-7}) further improves the translation quality of the MT-based systems to 73.1% in ABC, achieving a gain of 9.2% over conventional methods to select the best translation.

Table 10 Seed contribution of MT^r_{1-7}.

initial translation hypotheses	selected as translation (%)	translation accuracy ABC(%)
EBMT only (MT_{1-2})	60.1	68.5
OTSMT only (MT_{3-7})	37.4	79.3
MT_{1-2} and MT_{3-7}	2.5	83.3

4.3.3 Discussion

In this section, we discuss the experimental results obtained for the best system MT^r_{1-7} by (a) analyzing the contribution of EBMT-based and OTSMT-based decoder seeds, (b) analyzing the system performance for the source language perplexity of the input, and (c) comparing the overall system performance with the MT engines used to produce the initial translation hypotheses.

4.3.3.a Seed Contribution

Table 10 illustrates the percentage of the respective initial translation hypotheses of each type that were decoded into the selected translation of the best system MT^r_{1-7}. The results show that:

- OTSMT hypotheses are used to generate nearly 2/5 of the translations.
- The overlap in initial translation hypotheses that produce the same translation is small, thus showing *the complementary effect of combining EBMT-based and OTSMT-based hypotheses.*
- The quality of the decoder output generated from OTSMT-based hypotheses is

☆ For this experiment, we used up to 16 human reference translations.

Table 11 Translation accuracy (%) of selected translations.

initial translation hypotheses		source language perplexity LOW	MED	HIGH
EBMT only	MT_{1-2}	**88.9**	59.9	57.1
OTSMT only	MT_{3-7}	82.2	**82.9**	**72.4**

high. The percentage of selected OTSMT-based hypotheses that result in unacceptable translations (rank *D*) is only 20.7%, compared to 31.5% for the EBMT-based hypotheses. The translations produced by both types of initial translation hypotheses show the highest accuracy. Only 16.7% of the sentences are ranked as *D*.

4.3.3.b Analysis for Source Language Perplexity

In order to get an idea of how the EBMT-based and OTSMT-based hypotheses affect the overall system performance, we analyzed the dependency on the *source language perplexity* of the input sentence.

We calculated the language perplexity of the source language sentences of the test set and clustered them into three subsets of *low* (LOW), *medium* (MED), and *high* (HIGH) language perplexity ☆.

The percentage of selected translations ranked as *A*, *B*, or *C* for the three source language perplexity subsets are listed in **Table 11**.

The evaluation results show that:

- The EBMT-based hypotheses achieve better results for sentences of *low* language perplexity.
- The OTSMT-based hypotheses outperformed the EBMT-based hypotheses for sentences of *medium* and *high* language perplexity.

Therefore, the usage of all MT-based decoder seeds (MT^r_{1-7}) boosts the system performance by exploiting the strength of both types of initial hypotheses.

4.3.3.c Comparison with Utilized MT Engines

In this section, we compare the overall system performance of the proposed greedy decoder with the translation accuracy of the MT engines utilized to produce the initial translation hypotheses.

☆ The language perplexity thresholds were selected so that each subset consists of 1/3 of the test sentences, i.e., 19.4 (LOW) and 52.3 (HIGH).

Table 12 Comparison with utilized MT engines.

initial translation hypothesis		translation accuracy ABC (%)
EBMT	MT_1	21.5
	MT_2	67.7
OTSMT	MT_3	52.9
	MT_4	53.9
	MT_5	48.2
	MT_6	50.3
	MT_7	45.6
proposed greedy decoder (MT^r_{1-7})		**73.1**

The results listed in **Table 12** show that:

- The translation accuracy of the MT engines used to produce the initial translation hypotheses varies between 21.5% and 67.7%.
- *The proposed method outperforms all MT engines*, achieving a gain of 5.4% over the best translation engine MT_2.

4.3.4 Summary

Therefore, we can summarize the effects of using multiple MT-based decoder seeds in combination with the selection of the best translation according to the edit-distance-based rescoring method on greedy decoding as follows:

- *Multiple MT-based hypotheses can help to avoid local optima problems in the search* by exploiting large variations of translation engine architectures.
- *The proposed rescoring method drastically reduces problems in translation candidate selection solely based on statistical models* due to the incorporation of information on how much the initial translation hypotheses are modified during decoding.
- *The proposed method significantly outperforms the translation engines* used to produce the initial translation hypotheses.

4.4 Comparison to Related Research

In this section, we introduce previous counter-measures to the two problems addressed in this paper and compare them with the proposed methods.

The first problem of the greedy decoding approach was previously dealt with by applying example-based methods (cf. Section 3.1.1). These methods create an initial translation hypothesis based on translation examples, i.e., pairs of a source and a human-translated target language sentence, where the source sentence is *similar* to the given input.

- Marcu [11] extracted phrase translations to

fully cover the input sentence and uses the concatenation of the corresponding target phrases as the initial translation hypothesis.
- Watanabe and Sumita [22] utilized translation examples on the sentence level.

The advantage of these *example-based decoding* approaches is that the search for a good translation starts from a nearly correct hypothesis, if an appropriate translation example can be found. Section 4.4.1 compares the proposed MT-based method to the example-based decoder approach of Watanabe and Sumita [22].

The second problem of the greedy decoding approach is the selection of candidates of lower translation quality according to the statistical models. Both example-based decoding methods mentioned above try to avoid this problem by bypassing the decoding process itself, if a perfectly matching translation example is found in the parallel text corpus, i.e. the source part of the translation example matches the input sentence. In this case, the target part of the translation example is output as the translation.

Other researchers address the problem of how to select the best translation among multiple translation candidates by using multiple language and translation models.

- Nomoto [13] made use of voted language models to choose among outputs of multiple OTSMT engines. This method utilizes multiple language models trained on corpora of various genres ☆ with different vocabulary sizes. For a given input, it first selects the language model that gives the smallest target language perplexity for the majority of the given translation candidates. Next, it selects the best translation out of the translation candidates according to the maximal score obtained by the chosen model.
- Akiba, et al.[1] trained multiple language and translation model pairs from n-fold subsets of the training data to select the best translation candidate based on a multiple comparison test. This test checks whether the obtained $\mathrm{TM}{\cdot}\mathrm{LM}_{i=1,..,n}$ scores of one translation candidate are significantly higher than those of the others.

Section 4.4.2 compares the proposed rescoring method with the selector approach of Akiba, et al.[1]

☆ The genres used are *news articles*, *business-related text*, *patents*, and *literary texts*.

Table 13 Comparison with example-based decoding [22].

initial translation hypotheses		translation accuracy ABC (%)
EB	EB_{all}	53.9
	EB_{exact}	54.5
EBMT+	MT_{1-7}	63.9
OTSMT	$\mathbf{MT}^{\mathbf{r}}_{\mathbf{1-7}}$	**73.1**

4.4.1 Example-based Decoding

The example-based decoding approach of Watanabe and Sumita [22] (EB_{exact}) is an extension of the EB_{all} system described in Section 4.2.1. The initial translation hypotheses are retrieved from the training corpus, and the translation output is selected based on statistical scores only. However, the decoding process of seeds, whose source language input is identical to the source parts of translation examples in the training corpus (*exact match*), is skipped and the target language sentence of the retrieved translation example is used as the translation output.

The results given in **Table 13** show that:

- The MT_{1-7} system, which also selects translation candidates solely based on statistical models, outperforms both example-based methods. This indicates that *using multiple translation-engine-based initial translation hypotheses is more effective for overcoming the local optima problems in the search.*
- An additional improvement is obtained by the proposed rescoring method, achieving a gain of 18.6% in translation accuracy for MT^{r}_{1-7} over the example-based decoding approach EB_{exact}.
- Skipping the decoding process for all exactly matched translation examples retrieved from the training corpus achieves almost no improvement in translation accuracy. 15.5% of the test sentences were exact matches out of which only 4.6% were translations different from EB_{all}. Altogether, a gain of 0.6% in translation accuracy is achieved by EB_{exact} compared to EB_{all}.

4.4.2 Selection of Multiple Translation Engine Outputs

The selector approach of Akiba, et al.[1] utilizes multiple language and translation model pairs trained on different subsets of the training data to select the best translation among outputs from multiple MT engines. They use a

Table 14 Comparison with selection of multiple translation engine outputs [1].

initial translation hypotheses	translation accuracy ABC (%)
SEL	67.1
MT^{r}_{1-7}	**73.1**

multiple comparison test to check whether the obtained $TM{\cdot}LM_{i=1,..,n}$ scores of one MT output are significantly higher than those of the other MT outputs. If this is the case, that MT output is selected. Otherwise, the MT output with the highest average score is selected.

In order to compare the selector approach with the proposed rescoring method, we adopted the method of Akiba et al.[1] as follows:

- We randomly divided the training set into three subsets (S_i; $1 \leq i \leq 3$).
- We trained three different translation and language model pairs on all pairwise combinations of the subsets ($S_1 \cup S_2$, $S_1 \cup S_3$, $S_2 \cup S_3$).
- We applied the selector method (SEL) to select the best translation out of the seven MT-based single-seed decoder outputs $MT^{r}_{i=1,...,7}$. Therefore, both methods are applied to the same set of translation candidates.

In the case of the proposed method (MT^{r}_{1-7}), the translation candidate with the highest TM·LM·ED score is selected as the translation. The results given in **Table 14** show that:

- The rescoring method MT^{r}_{1-7} outperforms the selector approach SEL in terms of translation accuracy, gaining 6.0% in translation accuracy.
- *The compensation of the statistical scores* of each translation candidate by information on how much the initial translation hypothesis is modified during decoding *is more effective than applying multiple statistical models.*

5. Conclusions

This paper addressed two problems of conventional greedy decoding approaches for statistical machine translation, i.e., (1) the local optima problem due to the dependency on the initial translation hypothesis to start the search, and (2) the selection of candidates of lower translation quality solely based on statistical models.

We proposed two methods to overcome these problems by (1) using multiple translation-engine-based hypotheses to start the search and (2) selecting the best translation candidate by using an edit-distance-based rescoring method, which compensates the statistical scores of each translation candidate by using information on how much the initial translation hypothesis is modified during decoding. The proposed methods were integrated into the greedy decoding approach and the effectiveness of this approach was verified for Japanese-to-English translation of dialogues in the travel domain.

The proposed greedy decoding approach achieved a translation accuracy of 73.1%, which is an improvement of 18.6% over a previous greedy decoder [22] that relies solely on examples and statistical model scores as well as an improvement of 6.0% over a previous method [1] that uses multiple language and translation model pairs to select the best translation among multiple MT outputs. An analysis of the evaluation results showed that:

- *Multiple MT-based hypotheses can help to avoid local optima problems in the search* by exploiting large variations of translation engine architectures.
- *The proposed rescoring method drastically reduces problems in translation candidate selection solely based on statistical models* due to the incorporation of information on how much the initial translation hypotheses are modified during decoding.
- *The proposed method significantly outperforms the translation engines* used to produce the initial translation hypotheses.

References

1) Akiba, Y., Watanabe, T. and Sumita, E.: Using Language and Translation Models to Select the Best among Outputs from Multiple MT Systems, *Proc. COLING 2002*, Taipei, Taiwan, pp.8–14 (2002).
2) Brown, P., Pietra, S.D., Pietra, V.D. and Mercer, R.: The mathematics of statistical machine translation: Parameter estimation, *Computational Linguistics*, Vol.19, No.2, pp.263–311 (1993).
3) Cormen, H., Leiserson, C. and Rivest, L.: *Introduction to Algorithms*, MIT Press (1996).
4) Fujitsu: ATLAS Honyaku Superpack V9 (2003). http://software.fujitsu.com/jp/atlas
5) Germann, U., Jahr, M., Knight, K., Marcu, D. and Yamada, K.: Fast Decoding and Optimal Decoding for Machine Translation, *Proc. 39th*

ACL, Toulouse, France, pp.228–235 (2001).
6) IBM: Internet honyaku no oosama bilingual Version 5 (2001). http://www.ibm.com/jp/software/internet/king
7) Imamura, K.: Application of translation knowledge acquired by hierarchical phrase alignment for pattern-based MT, *Proc. 9th TMI*, Kyoto, Japan, pp.74–84 (2002).
8) Kikui, G., Sumita, E., Takezawa, T. and Yamamoto, S.: Creating Corpora for Speech-to-Speech Translation, *Proc. EUROSPEECH03*, Geneve, Switzerland, pp.381–384 (2003).
9) LogoVista: X PRO Multilingual Edition Ver.2.0 (2001). http://www.logovista.co.jp
10) Manning, C. and Schütze, H.: *Foundations of Statistical Natural Language Processing*, The MIT Press, Cambridge, Massachusetts (1999).
11) Marcu, D.: Towards a Unified Approach to Memory- and Statistical-Based Machine Translation, *Proc. 39th ACL*, Toulouse, France, pp.378–385 (2001).
12) NEC: CROSSROAD Ver3.0 (1999). http://meshplus.mesh.ne.jp/CROSSROAD/index.html
13) Nomoto, T.: Multi-Engine Machine Translation with Voted Language Model, *Proc. 42nd ACL*, Barcelona, Spain, pp.494–501 (2004).
14) Su, K., Wu, M. and Chang, J.: A New Quantitative Quality Measure for Machine Translation Systems, *Proc. 14th COLING*, Nantes, France, pp.433–439 (1992).
15) Sumita, E.: Example-based machine translation using DP-matching between word sequences, *Proc. 39th ACL, Workshop: Data-Driven Methods in Machine Translation*, Toulouse, France, pp.1–8 (2001).
16) Sumita, E., Yamada, S., Yamamoto, K., Paul, M., Kashioka, H., Ishikawa, K. and Shirai, S.: Solutions to Problems Inherent in Spoken-language Translation: The ATR-MATRIX Approach, *Proc. Machine Translation Summit VII*, Singapore, pp.229–235 (1999).
17) Takezawa, T., Sumita, E., Sugaya, F., Yamamoto, H. and Yamamoto, S.: Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world, *Proc. 3rd LREC*, Las Palmas, Spain, pp.147–152 (2002).
18) Tillmann, C. and Ney, H.: Word reordering and dp-based search in statistical machine translation, *Proc. COLING 2000*, Saarbruecken, Germany (2000).
19) Toshiba: The Honyaku Ver.7.0 (2003). http://pf.toshiba-sol.co.jp/prod/hon_yaku/index_j.htm
20) Wagner, R.: The string-to-string correction problem, *J. ACM*, Vol.21, No.1, pp.169–173 (1974).
21) Wang, Y. and Waibel, A.: Decoding algorithm in statistical machine translation, *Proc. 36th ACL*, Madrid, Spain (1997).
22) Watanabe, T. and Sumita, E.: Example-based Decoding for Statistical Machine Translation, *Proc. Machine Translation Summit IX*, New Orleans, USA, pp.410–417 (2003).

(Received May 11, 2005)
(Accepted September 2, 2005)
(Online version of this article can be found in the IPSJ Digital Courier, Vol.1, pp.561–575.)

Michael Paul received the B.E. and the M.E. degrees in computer science from the University of Saarland, in Saarbrücken/Germany, in 1992 and 1994 respectively. He is a research associate of the ATR Spoken Language Communication Research Laboratories. His research interests include natural language processing, (machine translation, evaluation), spoken language processing, and context processing. He is a member of the EAMT and the ACL.

Eiichiro Sumita received an M.S. degree in computer science from the University of Electro-Communications in 1982 and a Ph.D. degree in engineering from Kyoto University in 1999. His research interests include natural language processing (machine translation, paraphrasing and summarization), spoken language processing, information retrieval, e-Learning and parallel processing. He is currently leading a project on machine translation named Corpus-Centered Computation (C3) at ATR. He serves as Associate Editor of ACM Transactions on Speech and Language Processing. He is a member of the ACL, the IEICE, the IPSJ, the ASJ and the ANLP.

Seiichi Yamamoto was graduated from Osaka University in 1972 and received his Masters and Ph.D. degrees from Osaka University in 1974 and 1983, respectively. He joined Kokusai Denshin Denwa Co. Ltd. in April 1974, and ATR Interpreting Telecommunications Research Laboratories in May 1997. He was appointed president of ATR-ITL in 1997. He is professor of Doshisha University and invited researcher (ATR Fellow) at ATR Spoken Language Communication Research Laboratories at present. His research interests include digital signal processing, speech recognition, speech synthesis, natural language processing, and spoken language translation. He received Technology Development Awards from the Acoustical Society of Japan in 1995 and 1997. IEEE Fellow, Fellow of IEICE Japan.