A Machine Learning Approach to Hypotheses Selection of Greedy Decoding for SMT

Michael Paul and Eiichiro Sumita and Seiichi Yamamoto

ATR Spoken Language Communication Research Laboratories
2-2-2 Hikaridai, Kansai Science City, Kyoto, 619-0288 Japan
{michael.paul, eiichiro.sumita, seiichi.yamamoto}@atr.jp

Abstract

This paper proposes a method for integrating example-based and rule-based machine translation systems with statistical methods. It 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. In order to reduce *local optima* problems inherent in the search, the outputs generated by *multiple translation engines*, such as rule-based (RBMT) and example-based (EBMT) systems, are utilized as the initial translation hypotheses. This method outperforms conventional greedy decoding approaches using initial translation hypotheses based on translation examples retrieved from a parallel text corpus. However, the decoding of multiple initial translation hypotheses is computationally expensive. This paper proposes a method to select a single initial translation hypothesis *before decoding* based on a machine learning approach that judges the appropriateness of multiple initial translation hypotheses and selects the most confident *one* for decoding. Our approach is evaluated for the translation of dialogues in the travel domain, and the results show that it drastically reduces computational costs without a loss in translation quality.

1 Introduction

This paper proposes a method for integrating example-based and rule-based machine translation systems with statistical methods. It extends a greedy decoder for statistical machine translation (cf. Section 2), 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. Despite a high performance on average, the greedy decoding approach can often produce translations with severe errors.

A major problem of the greedy decoding approach is that 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. Therefore, the selection of the starting point is crucial to avoid local optima in the search.

Previous methods addressed this problem by creating an initial translation hypothesis based on translation examples obtained from a parallel text corpus (Marcu, 2001), (Watanabe and Sumita, 2003) or by using diverse starting points generated by multiple translation engines (Paul et al., 2004). 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. This method outperforms conventional greedy decoding approaches using initial translation hypotheses based on translation examples retrieved from a parallel text corpus. However, the sequential decoding of multiple decoder seeds is *computationally expensive.*

In this paper, we propose a method to select a single initial translation hypothesis *before decoding* in order to reduce computational costs. A machine learning approach (*decision tree*), that judges the appropriateness of a given initial translation hypothesis, is combined with a ranking method based on statistical model scores in order to select the most confident initial translation hypothesis for decoding. Section 3 extends the greedy decoding approach as follows: (1) the initial translation hypotheses are produced by multiple MT engines, (2) a machine learning approach using a decision tree classifier is proposed to identify and eliminate hypotheses that might be wrongly modified by the greedy decoder thus leading to translations of lower quality, and (3) information about the classification

117

result and statistical model scores of the remaining initial translation hypotheses are combined in order to select the best suited hypothesis.

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:

$$\arg\max_T \; p(S|T) * p(T), \qquad (1)$$

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 (Brown et al., 1993). During the translation 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 (Brown et al., 1993) 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* (Wang and Waibel, 1997), *beam search* (Tillmann and Ney, 2000), and *greedy decoding* (Germann et al., 2001). This paper concentrates on the greedy decoding approach described in details in Section 2.2.1. The local optima problem of this approach is illustrated in Section 2.2.2.
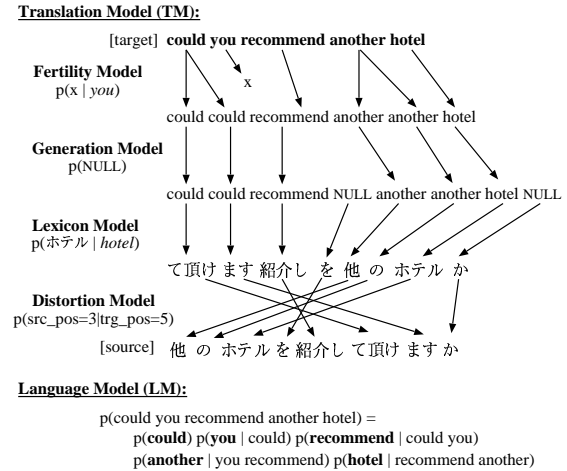
Translation Model (TM):

[target] **could you recommend another hotel**

Fertility Model
p(x | *you*)

x

could could recommend another another hotel

Generation Model
p(NULL)

could could recommend NULL another another hotel NULL

Lexicon Model
p(ホテル | *hotel*)

て頂けます 紹介し を 他 の ホテル か

Distortion Model
p(src_pos=3|trg_pos=5)

[source] 他 の ホテル を 紹介 して 頂け ます か

Language Model (LM):

p(could you recommend another hotel) =
    p(**could**) p(**you** | could) p(**recommend** | could you)
    p(**another** | you recommend) p(**hotel** | recommend another)

Figure 1: Statistical Models

2.2.1 Algorithm

Figure 2 illustrates the decoding algorithm, which is described in detail in (Germann et al., 2001), 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 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 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*.

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 Local Optima Problem of Greedy Decoding

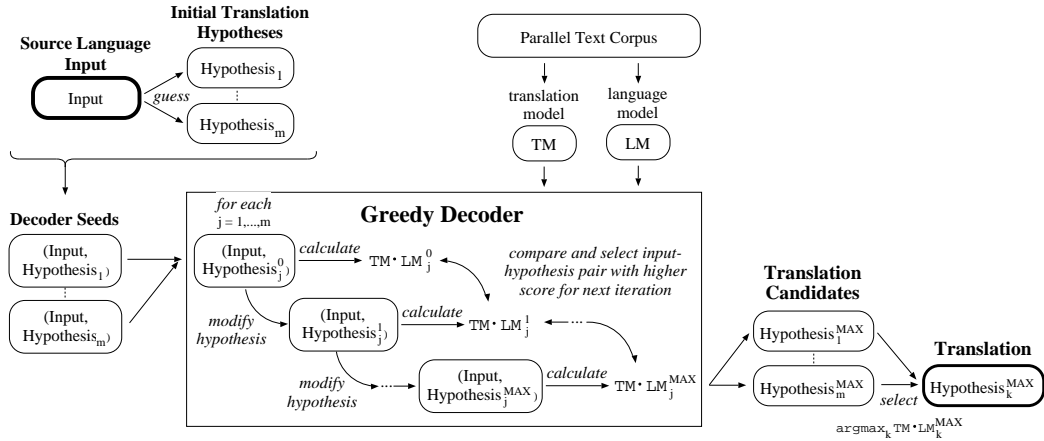A major problem of the greedy decoding approach is that the translation output depends

Figure 2: Greedy Decoding

on the initial translation hypothesis to start the search, which may lead to a local optimum translation but not to the global optimum translation.

This problem is illustrated in Figure 3. 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.
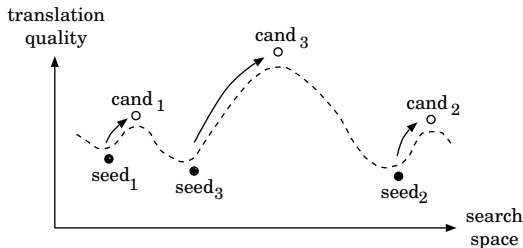

Figure 3: Local Optima Problem of the Greedy Search

2.3 Greedy Decoding Using Translation-Engine-Based Hypotheses

To solve the local optima problem, (Paul et al., 2004) proposed to use 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 obtained, because they are produced by independently developed translation engines that use different dictionaries, grammars, and translation rules. Therefore, larger parts of the search space can be explored, increasing the chance to catch the global optimum.

The greedy decoder is applied sequentially to each of the initial translation hypotheses, where the best translation is selected according to 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.

This method outperforms conventional greedy decoding approaches solely based on statistical models. However, a shortcoming of this approach is that the decoder has to be applied to *all* initial translation hypotheses. Therefore, high computational costs are involved to identify the best translation.

3 Machine Learning Approach for Hypotheses Selection

The method proposed in this paper is based on the greedy decoding approach described in Section 2.3. In order to reduce computational costs, our approach selects a single hypothesis out of the set of initial translation hypotheses obtained from multiple MT engines *before* the greedy decoder is applied to generate the translation output.

The initial translation hypotheses are produced by multiple MT engines as described in Section 3.1.

In order to select the most appropriate initial translation hypothesis for decoding, we propose a machine learning approach using a decision tree classifier to identify and eliminate hypotheses that might be wrongly modified by the

greedy decoder thus leading to translations of lower quality (cf. Section 3.2).

Finally, information about the classification result and statistical model scores of the remaining initial translation hypotheses are combined in order to select the best suited hypothesis as described in Section 3.3.

3.1 Translation-Engine-based Hypotheses

For our experiments, we used the five MT engines listed in Table 1[1].

Table 1: Utilized MT Engines

EBMT	D3 (Sumita, 2001)
	HPAT (Imamura, 2002)
RBMT	ATLAS (Fujitsu, 2003)
	LOGOVISTA (LogoVista, 2001)
	THEHONYAKU (Toshiba, 2003)

Two of them (MT_{1-2}) are *example-based MT* (EBMT) systems that are trained on the same training set as the greedy decoder. The remaining three (MT_{3-5}) are *off-the-shelf rule-based MT* (RBMT) systems that are based on lexicons, grammars, and translation rules. Examples of MT-based hypotheses are given in Table 2.

Table 2: Translation-Engine-Based Hypotheses

(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
($\rightarrow$ *i would prefer the hyatt regency please
and if possible i want a single room*)

(initial translation hypothesis)
MT_1: i 'm asked do i want to stay to room
single room
MT_2: i 'll send a hyatt 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

The outputs of each MT engine show large variations, because they are produced by independently developed translation engines that use different translation knowledge resources.

3.2 Decision Tree Classifier

We use a machine learning approach in order to learn an automatic decision tree classifier (Rulequest, 2004) that distinguishes between initial translation hypotheses being decoded into translations of low vs. high quality.

The decision tree classifier is trained on monolingual as well as bilingual features obtained for pairs of source language input sentences and MT engine outputs. The features were selected in order to cover inter-hypotheses characteristics as well as general features for the identification of appropriate initial translation hypotheses. The *inter-hypotheses features* consist of the following:

- **Similarity features** between initial translation hypotheses produced by different MT engines.

  - the number of identical initial translation hypotheses

  - the *average edit-distance* between the given hypotheses and those of other MT engines, whereby the edit-distance is defined as the sum of the costs of *insertion*, *deletion*, and *substitution* operations required to map one word sequence into the other (Wagner, 1974).

  - differences in the length of a given initial translation hypothesis toward the shortest/longest initial translation hypothesis.

Moreover, we added also *statistical features* and *syntactic/semantic features* for the experiments described in this paper, some of which were used in previous research on the automatic evaluation of machine translation output (Corston-Oliver et al., 2001).

- **Perplexity** of the source language input and the initial translation hypothesis calculated on the basis of trigram language models.

- **Translation model and language model scores** of the input-hypothesis pairs.

- **Dictionary features** including the number of OOV (out-of-vocabulary) words and the number of target words in the initial translation hypothesis that are possible translations of source words.

- **Syntactic features** that are extracted from the syntactic structure of the source language input and the initial translation hypotheses, respectively. These can be subcategorized as follows.

---

[1]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.

- *sentence length*

- *sentence type*

- *sentence parse* (success of parsing, number of nodes in the parse-tree, number/length of pre/post-modifiers of noun phrases, number of coordinated constituents, coordination balance, i.e., the maximal length difference in coordinated constituents)

- *size of constituents*

- *density features*, i.e., ratio of function words to content words

- **Semantic features** of content words that are extracted from a thesaurus (Ohno and Hamanishi, 1984).

During the *learning phase*, all MT engines listed in Table 1 are used to translate parts of the training corpus and to extract the above mentioned features automatically. Next, the greedy decoder is applied to each initial translation hypothesis, and the obtained results are evaluate automatically using the WER metrics introduced in Section 4.1.2. Based on this evaluation, each input-hypothesis pair is assigned to one of the following two classes:

$$class = \begin{cases} OK & \text{,if } WER(\textit{decoder output}) \\ & \quad < WER(\textit{initial translation} \\ & \quad\quad\quad \textit{hypothesis}) \\ NG & \text{,otherwise} \end{cases}$$

During the *application phase*, the obtained decision tree classifier is applied to each input-hypothesis pair. All initial translation hypotheses classified as *NG* are removed from the hypothesis set. In addition to the classification result, a *confidence score*, i.e., the percentage of training samples classified correctly using the same decision tree path, is assigned to each input-hypothesis pair.

3.3 Selection Algorithm

Statistical model scores are in general good indicators of translation quality and can be used to compare translation hypotheses directly. *The higher the statistical model score, the higher the translation quality is supposed to be.* However, the greedy decoding approach can often produce translations with severe errors. This occurs partly because the decoder might modify hypotheses wrongly resulting in translations of lower quality with higher statistical scores.

On the other hand, the decision tree classifier provides us with information about how reliable the decision is, i.e., *the higher the confidence score* derived from the classification result, *the more likely it is that a good starting point is found.* However, it is not possible to compare directly two hypotheses on the basis which one is more reliable than the other one, because the decision tree classifier is applied independently.

In order to select the most appropriate initial translation hypothesis classified as *OK*, we propose to use both types of information by combining the confidence score derived from the decision tree with the statistical model scores of the input-hypothesis pair (I, H) as follows:

$$\textbf{CONF·TM·LM}(I,H) = \frac{2*\text{conf}(I,H)*\text{logP}(TM·LM)}{\text{conf}(I,H)+\text{logP}(TM·LM)} ,$$

where $conf(I, H)$ is the confidence score derived from the classification result and *logP(TM·LM)* denotes the positive log-probabilities of the statistical model score for the given input-hypothesis pair (I, H).

The input-hypothesis pair with the highest CONF·TM·LM score is selected for decoding.

4 Evaluation

Section 4.1 describes the experimental setting. In order to train the translation[2] and language[3] models used for decoding, we utilize two corpora from the *travel* domain. The proposed method is evaluated by using an automatic evaluation metrics and a human assessment of *translation accuracy*. The baseline performance of the greedy decoder using multiple translation-engine-based hypotheses is given in Section 4.2. The effects of the hypotheses selection method proposed in this paper are summarized in Section 4.3 and the obtained results are discussed in Section 4.4.

4.1 Experimental Setting

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

4.1.1 Corpora

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

---

[2]The translation models are trained using the GIZA++ toolkit, http://www.fjoch.com

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

- *Basic Travel Expression Corpus* (BTEC)
  The BTEC corpus is a large collection of sentences[4] 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 (Takezawa et al., 2002).

- *Machine Aided Dialogue Corpus* (MAD)
  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 (Kikui et al., 2003).

The statistics of the corpora are given in Table 3, 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.

Table 3: Corpus Statistics

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

The BTEC corpus was used for the acquisition of translation knowledge (*training set*) and the MAD corpus was used for the training of the decision tree classifier. In addition, we used 502 sentences from the MAD corpus reserved for evaluation purposes as the test set.

4.1.2 Evaluation Metrics

For the evaluation, we used the following automatic scoring measure and human assessment.

- *Word Error Rate* (Su et al., 1992) (WER), which penalizes edit operations against reference translations..

- *Translation Accuracy* (Sumita et al., 1999) (ABC): subjective evaluation ranks ranging from A to D (A: perfect, B: fair, C: acceptable and D: nonsense), judged by a native speaker. Hereafter, we use the total count of translations ranked A, B, or C as the ABC score.

---

[4]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).

In contrast to WER, higher ABC scores indicate better translations. For the automatic scoring measure we utilized up to 16 human reference translations.

4.2 Translation-Engine-based Hypotheses

Table 4 summarizes the translation quality of the MT engines used to create the initial translation hypotheses.

Table 4: Utilized MT Engines

initial translation hypotheses		evaluation	
		WER (%)	ABC (%)
EBMT	MT_1	49.6	60.3
	MT_2	52.0	66.3
RBMT	MT_3	69.6	54.5
	MT_4	69.4	59.3
	MT_5	71.4	54.1

Table 5 summarizes the translation quality of the greedy decoder using the combination of all MT engine outputs as the initial translation hypotheses.

Table 5: Greedy Decoder Output

initial translation hypotheses	evaluation	
	WER (%)	ABC (%)
EBMT+RBMT (MT_{1-5})	45.8	67.7

The results demonstrate experimentally the effectiveness of using multiple translation-engine-based hypotheses for decoding. The greedy decoding approach (EBMT+RBMT) outperforms all MT engines used to create the initial hypotheses, gaining 3.8% in WER and 1.4% in ABC toward the best MT engine.

4.3 Hypotheses Selection Method

The translation of the MAD corpus by all MT engines listed in Table 1, resulted in 24,470 input-hypothesis pairs from which the feature sets described in Section 3.2 were extracted automatically. Based on this data set, a decision tree classifier was learned and its performance was evaluated as described in Section 4.3.1.

Next, the decision tree classifier was used to filter-out inappropriate initial translation hypotheses and the performance of the proposed selection method was evaluated as described in Section 4.3.2.

4.3.1 Performance of Decision Tree Classifier

Table 6 gives the percentage of sentences classified correctly (*actual = predicted*) and the total amount of classification errors for the training and test sentences, respectively.

Table 6: Decision Tree Classifier

(training corpus)

predicted	actual		
	OK	NG	total
OK	53.5	21.5	75.0
NG	6.9	18.1	25.0
total	60.4	39.6	

(test corpus)

predicted	actual		
	OK	NG	total
OK	66.5	14.5	81.0
NG	14.6	4.4	19.0
total	81.1	18.9	

In total, 72.6%/70.9% of the training/test set were classified correctly, where 21.5%/14.5% of the sentences were accepted falsely. However, 6.9%/14.6% of good initial translation hypotheses were ignored resulting in a total error of 18.4% for the training set and 29.1% for the test set.

4.3.2 Selection of Initial Translation Hypothesis

In order to investigate the effects of applying the decision tree classifier to the test sentences, we evaluated two different selection methods: the hypothesis with (1) the highest statistical score ($TM{\cdot}LM$), and (2) the highest CONF·TM·LM score is selected as the initial translation hypothesis to be used for decoding.

Table 7: Hypothesis Selection

selection method	evaluation	
	WER (%)	ABC (%)
TM·LM	53.2	61.1
CONF·TM·LM	48.1	67.9

The results summarized in Table 7 show, that:

- a large gain in performance is achieved for the combination of confidence scores with statistical model scores.

- the proposed method outperforms all single MT engines (cf. Table 4)

- it achieves the same level of performance as the sequential decoding of *all* initial translation hypotheses (cf. Table 5)

4.4 Discussion

In order to investigate the effects of the proposed method on the computational costs, we compared the processing time of the EBMT+RBMT system that decodes all five initial hypothesis toward the proposed CONF·TM·LM method that selects a single hypothesis. The results show that the proposed method is 7 times faster than the EBMT+RBMT system, thus reducing the computational costs by 85.7%.

Moreover, an investigation into the feature dependency revealed, that *inter-hypotheses* features are most important. For example, if two or more MT engines produce the same initial translation hypothesis, it is an indicator of good quality. Therefore, similarity features like "*the number of identical initial translation hypotheses*" appear at the top of the decision tree classifier.

On the other hand, *general features* like language perplexity or information about the sentence structure seems to be less important. They are used in the decision tree classifier, but appear mainly on lower levels of the decision tree.

However, the set of features used in our experiments is not exclusive. Further investigations have to verify the usefulness of additional features not used in the above experiments like the *minimal tiling of substrings* (Quirk, 2004).

Moreover, the lower total error rate obtained for the classification of the training compared to the test data set indicates the problem of overfitting. Therefore, the application of pruning techniques and the careful selection of features might help to improve the classifier performance and thus the overall system performance of the proposed method.

5 Conclusion

This paper described a machine learning approach to seeding a greedy decoder effectively. The proposed method used a decision tree classifier to judge the appropriateness of multiple translation-engine-based hypotheses and selects a single initial translation hypothesis *before* decoding based on statistical model scores of the input-hypothesis pairs as well as confidence scores derived from the decision tree classification results.

The proposed method was 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.

An analysis of the evaluation results showed that *the proposed hypotheses selection method avoids high computational costs* by limiting the decoding process to a single initial hypothesis *without a loss in translation quality.*

References

P. Brown, S. Della Pietra, V. Della Pietra, and R. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

S. Corston-Oliver, M. Gamon, and C. Brockett. 2001. A machine learning approach to the the automatic evaluation of machine translation. In *Proc. of 39th ACL*, pages 148–155, Toulouse, France.

Fujitsu. 2003. ATLAS Honyaku Superpack V9. http://software.fujitsu.com/jp/atlas.

U. Germann, M. Jahr, K. Knight, D. Marcu, and K. Yamada. 2001. Fast decoding and optimal decoding for machine translation. In *Proc. of 39th ACL*, pages 228–235, Toulouse, France.

K. Imamura. 2002. Application of translation knowledge acquired by hierarchical phrase alignment for pattern-based MT. In *Proc. of 9th TMI*, pages 74–84, Kyoto, Japan.

G. Kikui, E. Sumita, T. Takezawa, and S. Yamamoto. 2003. Creating corpora for speech-to-speech translation. In *Proc. of the EUROSPEECH03*, pages 381–384, Geneve, Switzerland.

LogoVista. 2001. X PRO Multilingual Edition Ver.2.0. http://www.logovista.co.jp.

D. Marcu. 2001. Towards a unified approach to memory- and statistical-based machine translation. In *Proc. of the 39th ACL*, pages 378–385, Toulouse, France.

S. Ohno and M. Hamanishi. 1984. *Ruigo-Shin-Jiten*. Kadokawa.

M. Paul, E. Sumita, and S. Yamamoto. 2004. Example-based rescoring of statistical machine translation output. In *Proc of the HLT-NAACL, Companion Volume*, pages 9–12, Boston, USA.

C.B. Quirk. 2004. Training a sentence-level machine translation confidence measure. In *Proc. of 4th LREC*, pages 825–828, Lisbon, Portugal.

Rulequest. 2004. Data mining tool c5.0. http://rulequest.com/see5-info.html.

K. Su, M. Wu, and J. Chang. 1992. A new quantitative quality measure for machine translation systems. In *Proc. of the 14th COLING*, pages 433–439, Nantes, France.

E. Sumita, S. Yamada, K. Yamamoto, M. Paul, H. Kashioka, K. Ishikawa, and S. Shirai. 1999. Solutions to problems inherent in spoken-language translation: The ATR-MATRIX approach. In *Proc. of the Machine Translation Summit VII*, pages 229–235, Singapore.

E. Sumita. 2001. Example-based machine translation using DP-matching between word sequences. In *Proc. of the 39th ACL, Workshop: Data-Driven Methods in Machine Translation*, pages 1–8, Toulouse, France.

T. Takezawa, E. Sumita, F. Sugaya, H. Yamamoto, and S. Yamamoto. 2002. Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. In *Proc. of the 3rd LREC*, pages 147–152, Las Palmas, Spain.

C. Tillmann and H. Ney. 2000. Word reordering and dp-based search in statistical machine translation. In *Proc. of COLING 2000*, Saarbruecken, Germany.

Toshiba. 2003. TheHonyaku Ver.7.0. http://pf.toshiba-sol.co.jp/prod/hon_yaku/index_j.htm.

R.W. Wagner. 1974. The string-to-string correction problem. *Journal of the ACM*, 21(1):169–173.

Y. Wang and A. Waibel. 1997. Decoding algorithm in statistical machine translation. In *Proc. of 36th ACL*, Madrid, Spain.

T. Watanabe and E. Sumita. 2003. Example-based decoding for statistical machine translation. In *Proc. of the Machine Translation Summit IX*, pages 410–417, New Orleans, USA.