Automatic machine translation selection scheme to output the best result

Keiji YASUDA*†, Fumiaki SUGAYA*, Toshiyuki TAKEZAWA*,
Seiichi YAMAMOTO*, Masuzo YANAGIDA†

* ATR Spoken Language Translation Research Laboratories
2-2-2 Hikari-dai Seika-cho, Soraku-gun, Kyoto, 619-0288, Japan
{keiji.yasuda, fumiaki.sugaya, toshiyuki.takezawa, seiichi.yamamoto}@atr.co.jp

†Doshisha University
1-3, Tatara-miyakodani, Kyotanabe, Kyoto, 610-0394, Japan
myanagid@mail.doshisha.ac.jp

Abstract

An automatic selection method for an integrated multiple MT system is proposed. This method employs a machine learning approach to build an automatic MT selector. The selector learns based on the parameters of MT systems and the evaluation result provided by a human evaluator. An experiment is conducted on two MT systems developed in our laboratories. Experimental results show the effectiveness of the proposed method. The ratio of correct selection is 76%. According to the system performance evaluation result, the integrated MT system using the proposed method gives a better performance than each individual MT system.

1. Introduction

No perfect Machine Translation (MT) system has yet appeared that can translate sentences for any kind of task/domain or form of expression. Some MT systems have an advantage in idiomatic phraseology, others have an advantage in non-idiomatic phraseology. We could construct an integrated MT system that performs better than each individual MT system, if we could use several MT systems complementarily.

Our laboratories have conducted many studies on MT systems. The scheme suggested here, however, has never been examined in our laboratories up to now.

In this paper, we propose an automatic selection method for an integrated multiple MT system and show experimental results for two MT systems. One of the MT systems is the Example Based Machine Translation (EBMT) system, which has been developed at ATR Spoken Language Translation Research Laboratories, and the other is the Transfer Driven Machine Translation (TDMT) system, which was developed at ATR Interpreting Telecommunications Research Laboratories.

In section 2, we briefly describe these MT systems. In section 3, we show an experimental result and performance evaluation of the automatic selector's and MT systems. In section 5 we state our conclusion.

2. Description of MT systems

EBMT (Sumita, 2001) and TDMT (Sumita et al., 1999) employ completely different translation strategies, and they have features that differ from each other. It is, then, likely that we can use these MT systems complementarily.

In this section, brief explanation is given about these MT systems.

2.1. EBMT

EBMT employs a simple translation strategy, and it requires only morphological analysis, but not parsing. In this method, word-to-word DP matching is carried out to retrieve an example in a parallel corpus. The retrieval is made based on DP-distance. The DP-distance between an input sentence and each sentence in a source language corpus can be calculated as follows:

$$D_{DP} = \frac{I + D + 2\sum D_s}{L_{input} + L_{example}} \quad (1)$$

where D_{DP} is the DP-distance, L_{input} is the total number of words in an input sentence, $L_{example}$ is the total number of words in a sentence in the corpus, I is the number of inserted words comparing an input sentence to a sentence in the corpus, D is the number of deleted words comparing an input sentence to a sentence in the corpus, and D_S is the semantic distance between a word in an input sentence and a word in a sentence in the corpus.

DP-distance indicates the semantic distance between an input sentence and each sentence in a source language corpus. This method uses only the example from the corpus having the shortest distance to the input sentence. Finally, it substitutes words in the example in the target language to yield a translation result.

2.2. TDMT

In TDMT, translation is mainly performed by a transfer process that applies a piece of transfer knowledge about the language-pair to an input sentence.

TDMT also uses a particular parsing method to deal with ill-formed input sentences. This method splits the ill-formed input into well-balanced translation units. The ill-formed sentence is split based on a score (syntax structure score) for the substructure.

The complete translation result is formed by concatenating the partial translation results of split units. The syntax structure score indicates how semantically similar the input sentence is to a set of pieces of applied translation knowledge, and how acceptable the form of the input sentence is.

2.3. Feature of the MT systems

Since TDMT uses pieces of transfer knowledge extracted from an entire parallel corpus, it is robust against expression differences between learning data and

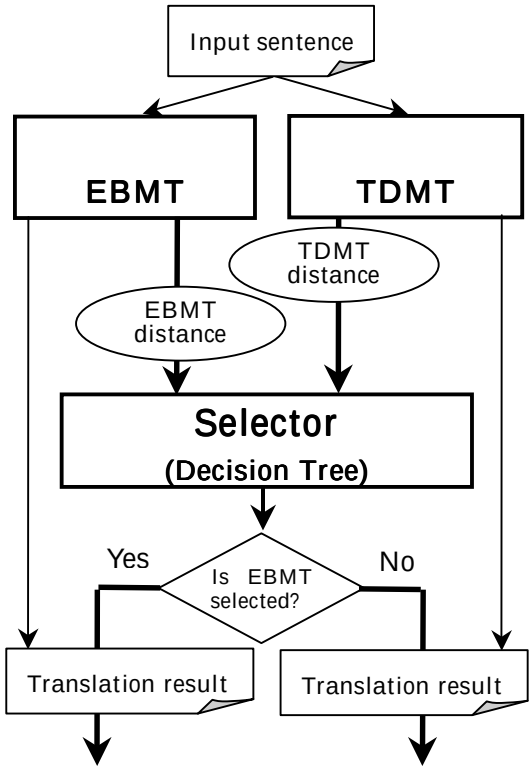

Figure 1: A diagram of the usage of the selector

input data. EBMT, on the other hand, is not as robust as TDMT, because it uses only the best-matched example translation pair, which it retrieves from the parallel corpus. Nevertheless, EBMT can yield high-quality translations in case an input sentence approximately coincides with a sentence in the learning corpus.

3. Integration method

We propose an automatic selection method for an integrated multiple MT system. In this paper, we employ the two MT systems described in the previous section complementarily. To make an automatic selector, we use the decision tree learner C5.0 (RULEQUEST RESEARCH, 2002), which is well known in the machine learning community.

3.1. Data configuration

Inputs to the selector are the syntax structure score from TDMT and the DP-distance from EBMT. We use the term "EBMT distance" to refer to DP-distance, and "TDMT distance" to refer to syntax structure score.

The selector to be made is a binary selector, which decides whether the TDMT output or the EBMT output is dominant. Figure 1 shows a diagram of the usage of the selector.

In a learning process, we use EBMT distance, TDMT distance, and the evaluation result determined by a human evaluator. The evaluation result is regarded as a teacher signal to teach the decision tree. Figure 2 shows a diagram of the learning process. In this process, the human evaluator selects the best translation for each learning sentence. If the evaluator is not able to select the best translation, i.e., both of the MT systems output the same quality translation, the learning sentence is rejected and will not be used for learning. Table 1 shows examples of the input data to the decision tree learner.

4. Experiment

Figure 3 shows details of a learning set and a test set. In this figure, the ordinate indicates the ratio to whole learning set or test set. The white area stands for "EBMT selected", which means that the human evaluator selected the EBMT output as the best translation. The black area stands for "TDMT selected". The gray area stands for "EVEN", which means both of the MT systems output translation of the same quality. Each number in the bar area indicates the number of sentences. As shown in Figure 3, the original learning set consists of 508 sentences, and 144 sentences are evaluated as "EVEN". So the actual size of the learning set used in the current study is 364 sentences. As shown in Figure 3, the test set also contains sentences, which generate "EVEN" results. We conduct two evaluations on these experiments. The first is focused on the selector's performance, and the second is focused on each of the MT systems' performance. The first evaluation is done on the "EVEN" excluded test set, consisting of 375 sentences, and the second evaluation, on the "EVEN" included test set, consisting of 510 sentences.

4.1. Experimental result

Figure 4 shows the distribution of the test set on the 2-dimensional (TDMT distance-EBMT distance) space. In this figure, the abscissa represents the EBMT distance, and the ordinate, the TDMT distance. Blank circles indicate EBMT-selected test sentences, and filled circles indicate TDMT-selected test sentences. The broken line on the figure is a learned boundary for the automatic selection. Upper left from the boundary is the EBMT portion and lower right from the boundary is the TDMT

System Parameter		Teacher Signal
EBMT Distance	TDMT Distance	Evaluation Result by Human Evaluator
1	4.57	TDMT
0	0	TDMT
0.2	25	EBMT
0.2	0.67	TDMT
0	0.83	EBMT
⋮	⋮	⋮
0	1.33	EBMT

Table 1: Examples of the input data to the decision tree learner

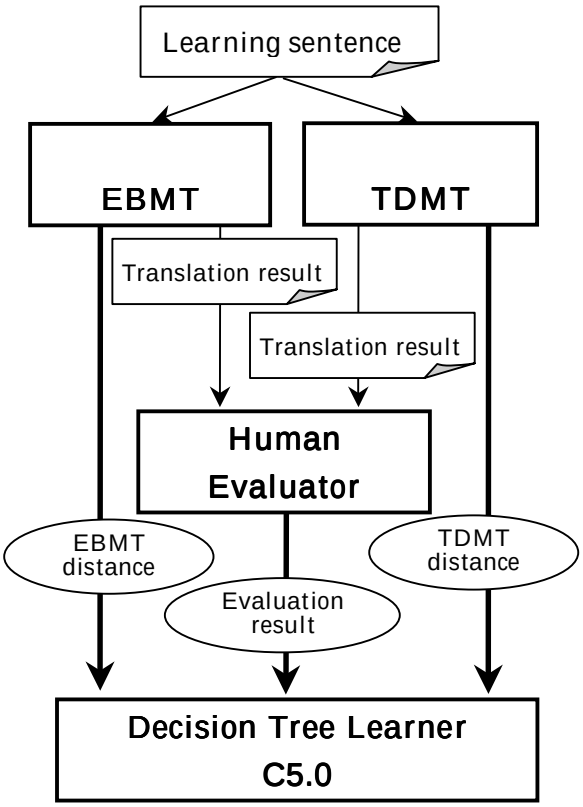

Figure 2: A diagram of the learning process

portion selected by the automatic selector. This boundary is learned on the learning set, using the decision tree learner.

4.2. Evaluation of the selector's performance

As mentioned before, the evaluation of the automatic selector is carried out for the "EVEN" excluded test set. For this evaluation, the selection result of the automatic selector is compared with the human evaluator's result. Table 2 shows the evaluation result. In the table, each number is an actual test sentence number, and each number in parentheses is the ratio to the whole "EVEN" excluded test set. The underlined number is the correctly selected result. As shown in this table, 76% of the test set is correctly selected.

4.3. Evaluation of the MT systems' performance

If the selector did not make any incorrect selections, an integrated MT system using an automatic selector would not give an inferior performance compared to each MT system. But, as shown in previous subsection, the selector has a 24% incorrect selection rate. In other words, there is no proof that the integrated system is superior to each MT system.

To make the effectiveness of the selector certain, we conducted an evaluation using the "translation paired comparison method" (Sugaya et al., 2000), developed in our laboratories. In this method, a human evaluator compares each MT system's translation results to the translation results of human subjects having various "Test of English for International Communication" (TOEIC, 2002) scores. Generally, the TOEIC score, which ranges from 10 to 990, is used as a measure of human speech translation capability.

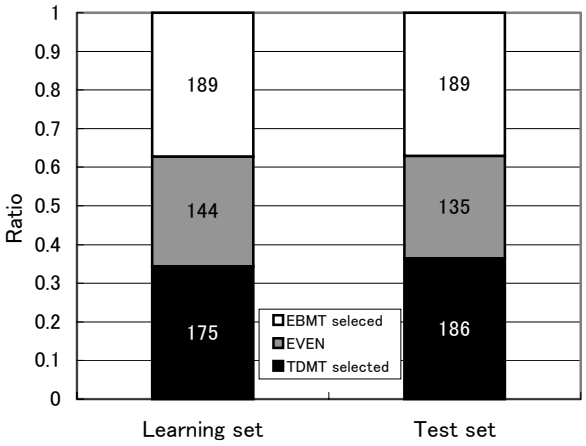

Figure 3: Detail of the data

In a conventional subjective evaluation method, such as a rank evaluation method, the ratio of sentences in each rank may be incapable of comparing each MT system's performance.

The main purpose of the translation paired comparison method is to estimate MT system's TOEIC capability as a TOEIC score. In this paper, we show the evaluation result comparing translation by the MT systems to translation by the subject with a 685 TOEIC score. This result is sufficient to demonstrate the effectiveness of the proposed method, although it is not sufficient to estimate the MT systems' TOEIC scores. To estimate the MT systems' TOEIC scores, we have to conduct an evaluation comparing translation by the MT systems to translation by several human subjects having various TOEIC scores.

As mentioned before, performance evaluation of the MT systems is carried out on the "EVEN" included test set. Figure 5 shows the evaluation result. In the figure, each number in the bar area indicates the number of sentences. The white area stands for "Human won", which means the translation by the human is better than that of MT. The black area stands for "MT won", which means the translation by MT is better than that by the human. The gray area stands for "Even". "Even" in this figure has a different meaning from "EVEN" in Figure 3. "Even" in Figure 5 means the translation by the MT system has the same quality as that by the human. Each arrow in the figure indicates the System Winning Rate (SWR) of each MT system. SWR can be calculated as follows:

$$SWR = \frac{W + 0.5 \times E}{T} \quad (2)$$

where T denotes the total number of sentences in the test set, W represents the number of "MT won" sentences, and E, the number of "Even" sentences. SWR signifies the degree capability of the MT system relative to that of the human.

As shown in the figure, the SWR of the integrated MT system using the proposed method is greater than that of TDMT or EBMT. This result shows the effectiveness of the proposed method.

The bar on the rightmost side shows the result of the integrated system using selection by the human evaluator, i.e., a perfect selector, who would not make incorrect selection. Considering this result, it is obvious that the EBMT and TDMT are complementary to each other, however, we still admit there is room for improvement of

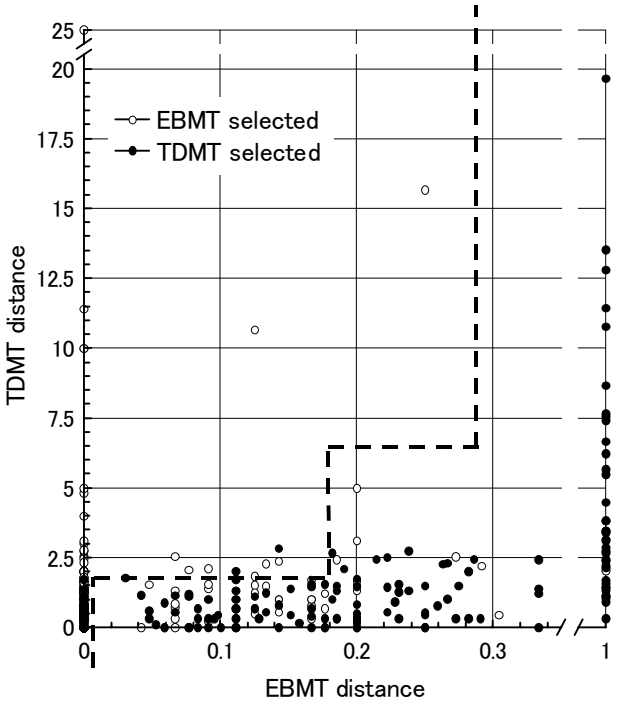

Figure 4: Distribution of the test sentences and learned boundary for the automatic selection

the automatic selection method in order to obtain more advantage.

Let us now discuss the subject from the point of view of improvement of the automatic selector. Looking at Figure 4, there is a limit to the improvement achieved by using parameters of the two MT systems because there are quite a few overlaps between the filled circles (for TDMT) and the blank circles (for EBMT). Taking this point into consideration, other parameters are needed to achieve an improvement, rather than a more efficacious machine learning algorithm. Hence, there is room for further investigation on necessity.

5. Conclusion

We proposed an automatic selection method for an integrated multiple MT system. In this method, the parameters of the MT systems and the evaluation results provided by a human evaluator are utilized to build an automatic selector using machine learning.

Two MT systems, EBMT and TDMT, are employed in the experiment. We conducted two evaluations. One was the evaluation of the automatic selector's performances, and the other was performance evaluation of the MT systems. According to the evaluation results, the ratio of correct selection was 76%, and the integrated system using the proposed method gave a better performance than each individual MT system.

In the near future, we will compare the performance of the proposed method with the performances of conventional methods (Callison-Burch et al., 2001; Tidhar et al., 2000). We will also carry out detailed evaluation to estimate TOEIC score of each MT systems.

Since our laboratories have been developing another MT system: Statistical Machine Translation (Watanabe et al., 2002), we are planning to integrate this system with the other two MT systems with the aim of further improving MT results.

		Evaluator's Selection	
		EBMT	TDMT
Decision Tree	EBMT	**135(36%)**	36(9.6%)
	TDMT	54(14.4%)	**150(40%)**

Table 2: Evaluation result of the selector

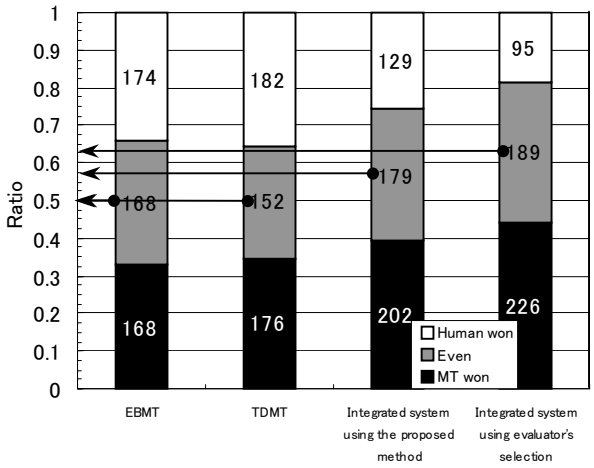

Figure 5: Evaluation result: comparison between each MT system's translation and translation by human with 685 TOEIC score

6. Acknowledgements

This research was supported in part by the Telecommunications Advancement Organization of Japan. It was also supported in part by the Academic Frontier Project promoted by Doshisha University.

7. References

Callison-Burch, C. and Flournoy, R.S., 2001. A Program for Automatic Selecting the Best Output from Multiple Machine Translation Engines, *Proc. MT Summit*, 63-66.

RULEQUEST RESEARCH, http://www.rulequest.com/

Sugaya, F., Takezawa, T., Yokoo, A., Sagisaka, Y. and Yamamoto, S., 2000. Evaluation of the ATR-MATRIX speech translation system with pair comparison method between the system and humans, *Proc. ICSLP*, 1105-1108.

Sumita, E., 2001. Example-based machine translation using DP-matching between word sequence, *Proc. ACL-2001 Workshop on Data-Driven Methods in Machine Translation*, 1-8.

Sumita, E., Yamada, S., Yamamoto, K., Paul, M., Kashioka, H., Ishikawa, K. and Shirai, S., 1999. Solutions to problems Inherent in spoken-language translation: The ATR-MATRIX approach, *Proc. MT Summit*, 229-235.

Tidhar, D. and Kussner, U., 2000. Learning to Select a Good Translation, *Proc.COLING2000*, pp.843-849.

TOEIC, http://www.toeic.com/

Watanabe, T., Imamura, K. and Sumita, E., 2002. Statistical Machine Translation Based on Hierarchical Phrase Alignment, *Proc. TMI 2002*, 188-198.