Example-based Machine Translation Based on Syntactic Transfer with Statistical Models

Kenji Imamura, Hideo Okuma, Taro Watanabe, and Eiichiro Sumita
ATR Spoken Language Translation Research Laboratories
2-2-2 Hikaridai, "Keihanna Science City"
Kyoto, 619-0288, Japan
{kenji.imamura,hideo.okuma,taro.watanabe,eiichiro.sumita}@atr.jp

Abstract

This paper presents example-based machine translation (MT) based on syntactic transfer, which selects the best translation by using models of statistical machine translation. Example-based MT sometimes generates invalid translations because it selects similar examples to the input sentence based only on source language similarity. The method proposed in this paper selects the best translation by using a language model and a translation model in the same manner as statistical MT, and it can improve MT quality over that of 'pure' example-based MT. A feature of this method is that the statistical models are applied after word re-ordering is achieved by syntactic transfer. This implies that MT quality is maintained even when we only apply a lexicon model as the translation model. In addition, translation speed is improved by bottom-up generation, which utilizes the tree structure that is output from the syntactic transfer.

1 Introduction

In response to the ongoing expansion of bilingual corpora, many machine translation (MT) methods have been proposed that automatically acquire their knowledge or models from the corpora. Recently, two major approaches to such machine translation have emerged: example-based machine translation and statistical machine translation.

Example-based MT (Nagao, 1984) regards a bilingual corpus as a database and retrieves examples that are similar to an input sentence. Then, a translation is generated by modifying the target part of the examples while referring to translation dictionaries. Most example-based MT systems employ phrases or sentences as the unit for examples, so they can translate while considering case relations or idiomatic expressions. However, when some examples conflict during re-

$E =$ NULL$_0$ show$_1$ me$_2$ the$_3$ one$_4$ in$_5$ the$_6$ window$_7$

$J =$ uindo$_1$ no$_2$ shinamono$_3$ o$_4$ mise$_5$ telidasai$_6$
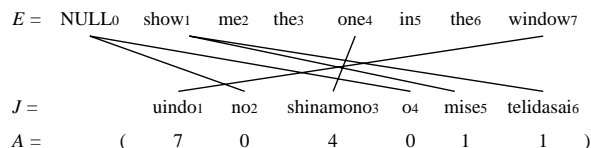
$A =$ (7 0 4 0 1 1)

Figure 1: Example of Word Alignment between English and Japanese (Watanabe and Sumita, 2003)

trieval, example-based MT selects the best example scored by the similarity between the input and the source part of the example. This implies that example-based MT does not check whether the translation of the given input sentence is correct or not.

On the other hand, statistical MT employing IBM models (Brown et al., 1993) translates an input sentence by the combination of word transfer and word re-ordering. Therefore, when it is applied to a language pair in which the word order is quite different (e.g., English and Japanese, Figure 1), it becomes difficult to find a globally optimal solution due to the enormous search space (Watanabe and Sumita, 2003).

Statistical MT could generate high-quality translations if it succeeded in finding a globally optimal solution. Therefore, the models employed by statistical MT are superior indicators of the quality of machine translation. Using this feature, Akiba et al. (2002) achieved selection of the best translation among those output by multiple MT engines.

This paper presents an example-based MT method based on syntactic transfer, which selects the best translation by using models of statistical MT. This method is roughly structured using two modules (Figure 2). One is an example-based syntactic transfer module. This module constructs
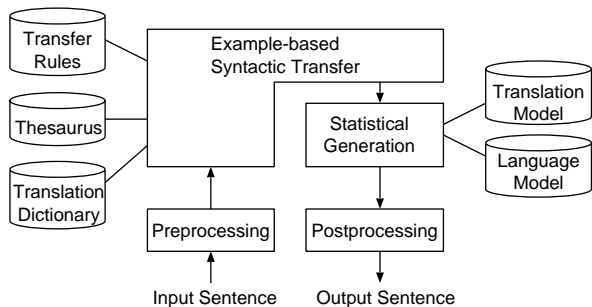
Figure 2: Structure of Proposed Method

tree structures of the target language by parsing and mapping the input sentence while referring to transfer rules. The other is a statistical generation module, which selects the best word sequence of the target language in the same manner as statistical MT. Therefore, this method is sequentially combined example-based and statistical MT.

The proposed method has the following advantages.

- From the viewpoint of example-based MT, the quality of machine translation improves by selecting the best translation not only from the similarity judgment between the input sentence and the source part of the examples but also from the scoring of translation correctness represented by the word transfer and word connection.

- From the viewpoint of statistical MT, an appropriate translation can be obtained even if we use simple models because a global search is applied after word re-ordering by syntactic transfer. In addition, the search space becomes smaller because the example-based transfer generates syntactically correct candidates for the most appropriate translation.

The rest of this paper is organized as follows: Section 2 describes the example-based syntactic transfer, Section 3 describes the statistical generation, Section 4 evaluates an experimental system that uses this method, and Section 5 compares other hybrid methods of example-based and statistical MT.

2 Example-based Syntactic Transfer

The example-based syntactic transfer used in this paper is a revised version of the Hierarchical Phrase Alignment-based Translator (HPAT, refer to (Imamura, 2002)). This section gives an overview with an example of Japanese-to-English machine translation.

2.1 Transfer Rules

Transfer rules are automatically acquired from bilingual corpora by using hierarchical phrase alignment (HPA; (Imamura, 2001)). HPA parses bilingual sentences and acquires corresponding syntactic nodes of the source and target sentences. The transfer rules are created from their node correspondences. Figure 3 shows an example of the transfer rules. Variables, such as X and Y in Figure 3, denote non-terminal symbols that correspond between source and target grammar. The set of transfer rules is regarded as synchronized context-free grammar.

The difference between this approach and conventional synchronized context-free grammar is that source examples are added to each transfer rule. The source example is an instance (i.e., a headword) of the variables that appeared in the training corpora. For example, the source example of Rule 1 in Figure 3 is obtained from a phrase pair of the Japanese verb phrase "*furaito* (flight) *wo yoyaku-suru* (reserve)" and the English verb phrase "*make a reservation for the flight.*"

2.2 Syntactic Transfer Process

When an input sentence is given, the target tree structure is constructed in the following three steps.

1. The input sentence is parsed by using the source grammar of the transfer rules.

2. The nodes in the source tree are mapped to the target nodes by using transfer rules.

3. If non-terminal symbols remain in the leaves of the target tree, candidates of translated words are inserted by referring to the translation dictionary.

An example of the syntactic transfer process is shown in Figure 4 for the input sentence "*basu wa 11 ji ni de masu* (The bus will leave at 11 o'clock)." There are two points worthy of notice in this figure. First, nodes in which the word order is inverted are generated after transfer (cf. VP node represented by a bold frame). Word re-ordering is achieved by syntactic transfer. Second, words

No.	Source Grammar		Target Grammar	Source Example
1	$VP \rightarrow X_{PP} Y_{VP}$	$\Rightarrow$	$VP \rightarrow Y_{VP} X_{PP}$	((*furaito* (flight), *yoyaku-suru* (reserve)) ..)
2			$VP \rightarrow Y_{VP} X_{ADVP}$	((*soko* (there), *yuku* (go)) ..)
3			$VP \rightarrow Y_{BEVP} X_{NP}$	((*hashi* (bridge), *aru* (be)) ..)
4	$S \rightarrow X_{NP}$ *wa* Y_{VP} *masu*	$\Rightarrow$	$S \rightarrow X_{NP} Y_{VP}$	((*kare* (he), *enso-suru* (play)) ..)
5			$S \rightarrow X_{NP}$ *will* Y_{VP}	((*basu* (bus), *tomaru* (stop)) ..)

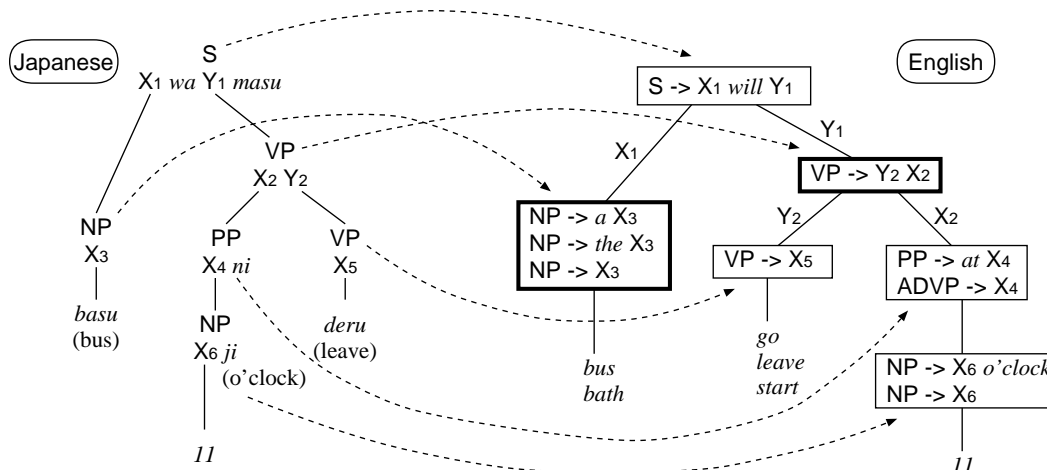Figure 3: Example of Transfer Rules

Figure 4: Example of Syntactic Transfer Process
(Bold frames are syntactic nodes mentioned in text)

that do not correspond between the source and target sentences (e.g., the determiner '*a*' or '*the*') are automatically inserted or eliminated by the target grammar (cf. NP node represented by a bold frame). Namely, transfer rules work in a manner similar to the functions of distortion, fertility, and NULL in IBM models.

2.3 Usage of Source Examples

Example-based transfer utilizes the source examples for disambiguation of mapping and parsing. Specifically, the semantic distance (Sumita and Iida, 1991) is calculated between the source examples and the headwords of the input sentence, and the transfer rules that contain the nearest example are used to construct the target tree structure. The semantic distance between words is defined as the distance from the leaf node to the most specific common abstraction (MSCA) in a thesaurus (Ohno and Hamanishi, 1984).

For example, if the input phrase "*ie* (home) *ni kaeru* (return)" is given, Rules 1 to 3 in Figure 3 are used for the syntactic transfer, and three target

nodes are generated without any disambiguation. However, when we compare the source examples with the headword of the variables X (*ie*) and Y (*kaeru*), only Rule 2 is used for the transfer because the semantic distance of the example (*soko* (there), *yuku* (go)) is the nearest. In the current implementation, all rules that contain examples of the same distance are used.

Consequently, example-based transfer achieves translation while considering case relations or idiomatic expressions based on the semantic distance from the source examples.

3 Statistical Generation

3.1 Translation Model and Language Model

Statistical generation searches for the most appropriate sequence of target words from the target tree output from the example-based syntactic transfer. The most appropriate sequence is determined from the product of the translation model and the language model in the same manner as statistical MT. In other words, when F and E denote the channel target and channel source sequence,

respectively, the output word sequence $\hat{E}$ that satisfies the following equation is searched for.

$$\begin{aligned} \hat{E} &= \underset{E}{\operatorname{argmax}} P(E|F) \\ &= \underset{E}{\operatorname{argmax}} P(E)P(F|E). \end{aligned} \quad (1)$$

We only utilize the lexicon model as the translation model in this paper, similar to the models proposed by Vogel et al. (2003). Namely, when f and e denote the channel target and channel source word, respectively, the translation probability is computed by the following equation.

$$P(F|E) = \prod_j \sum_i t(f_j|e_i). \quad (2)$$

The IBM models include other models, such as fertility, NULL, and distortion models. As we described in Section 2.2, the quality of machine translation is maintained using only the lexicon model because syntactical correctness is already preserved by example-based transfer.

For the language model, we utilize a standard word n-gram model.

3.2 Bottom-up Generation

We can construct word graphs by serializing the target tree structure, which allows us to select the best word sequence from the graphs. However, the tree structure already shares nodes transferred from the same input sub-sequence. The cost of calculating probabilities is equivalent if we calculate the probabilities while serializing the tree structure. We call this method bottom-up generation in this paper.

Figure 5 shows a partial example of bottom-up generation when the target tree in Figure 4 is given. For each node, word sub-sequences and their probabilities (language and translation) are obtained from child nodes. Then, the new probabilities of the word sequence combination are calculated, and the n-best sequences are selected. These n-best sequences and their probabilities are reused to calculate the probabilities of parent nodes. When the translation probability is calculated, the source word sub-sequence is obtained by tracing transfer mapping, and the applied translation model is restricted to the source sub-sequence. In other words, the translation probability is locally calculated between the corresponding phrases.

Set Name	Item	English	Japanese
Training	# of Sentences	152,170	
	# of Words	886,708	1,007,484
Test	# of Sentences	510	
	# of Words	2,973	3,340

Table 1: Corpus Size

When the generation reaches the top node, the language probability is re-calculated with marks for start-of-sentence and end-of-sentence, and the n-best list is re-sorted. As a result, the translation "*The bus will leave at 11 o'clock*" is obtained from the tree of Figure 4.

Bottom-up generation calculates the probabilities of shared nodes only once, so it effectively uses tree information.

4 Evaluation

In order to evaluate the effect when models of statistical MT are integrated into example-based MT, we compared various methods that changed the statistical generation module.

4.1 Experimental Setting

Bilingual Corpus   The corpus used in the following experiments is the Basic Travel Expression Corpus (Takezawa et al., 2002; Kikui et al., 2003). This is a collection of Japanese sentences and their English translations based on expressions that are usually found in phrasebooks for foreign tourists. We divided it into subsets for training and testing as shown in Table 1.

Transfer Rules   Transfer rules were acquired from the training set using hierarchical phrase alignment, and low-frequency rules that appeared less than twice were removed. The number of rules was 24,310.

Translation Model and Language Model   We used a lexicon model of IBM Model 4 learned by GIZA++ (Och and Ney, 2003) and word bigram and trigram models learned by CMU-Cambridge Statistical Language Modeling Toolkit (Clarkson and Rosenfeld, 1997).

Compared Methods   We compared the following four methods.

- **Baseline** (Example-based Transfer only)
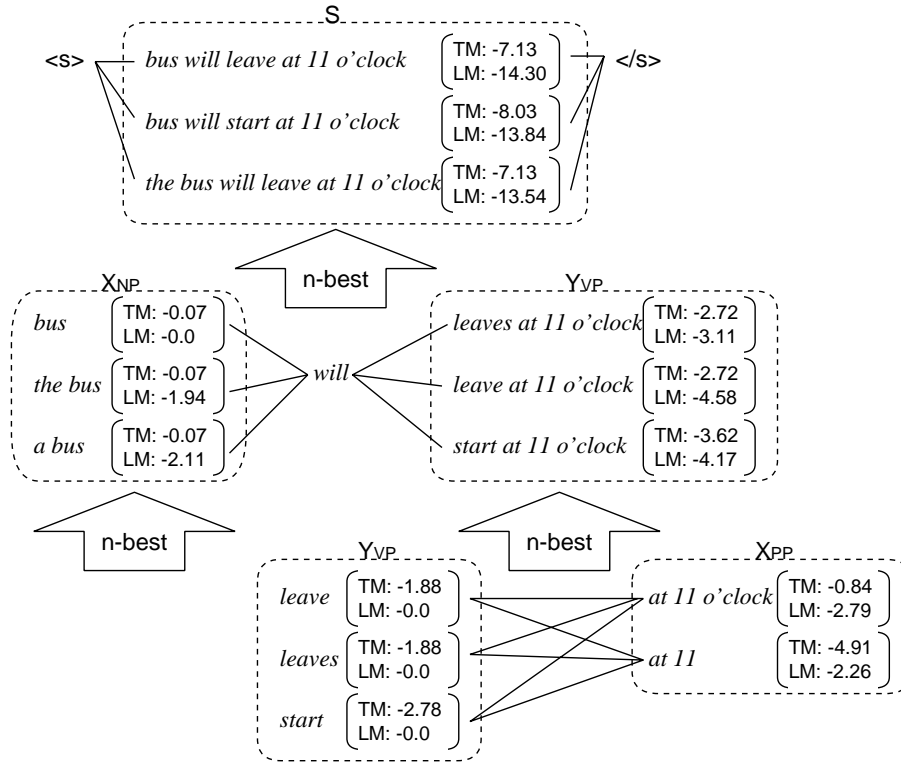 The best translation that had the same semantic distance was randomly selected from the

Figure 5: Example of Bottom-up Generation
(TM and LM denote log probabilities of the translation and language models, respectively)

tree that was output from the example-based transfer module. The translation words were selected in advance as those having the highest frequency in the training corpus. This is the baseline for translating a sentence when using only the example-based transfer.

- **Bottom-up**
  The bottom-up generation selects the best translation from the outputs of the example-based transfer. We used a 100-best criterion in this experiment.

- **All Search**
  For all combinations that can be generated from the outputs of the example-based transfer, we calculated the translation and language probabilities and selected the best translation. Namely, a globally optimal solution was selected when the search space was restricted by the example-based transfer.

- **LM Only**
  In the same way as All Search, the best translation was searched for, but only the language

model was used for calculating probabilities. The purpose of this experiment is to measure the influence of the translation model.

Evaluation Metrics From the test set, 510 sentences were evaluated by the following automatic and subjective evaluation metrics. The number of reference translations for automatic evaluation was 16 per sentence.

  BLEU: Automatic evaluation by BLEU score (Papineni et al., 2002).

  NIST: Automatic evaluation by NIST score (Doddington, 2002).

  mWER: The mean rate by calculating the word error rates between the MT results and all reference translations, where the lowest rate is selected.

  Subjective Evaluation: Subjective evaluation by an English native speaker into the four ranks of A: Perfect, B: Fair, C: Acceptable, and D: Nonsense.

Method	Automatic Evaluation			Subjective Evaluation			Translation Speed	
	BLEU	NIST	mWER	A	A+B	A+B+C	Mean (sec./sent.)	Worst (sec.)
Baseline	0.410	9.06	0.423	51.6%	64.3%	70.4%	0.180	10.82
Bottom-up	0.491	9.99	0.366	62.2%	72.5%	80.4%	0.211	5.03
All Search	0.498	10.04	0.353	62.9%	73.1%	80.8%	1.23	171.31
LM Only	0.491	9.11	0.385	57.6%	66.9%	72.0%	1.624	220.69

Table 2: MT Quality and Translation Speed vs. Generation Methods

4.2 Results

Table 2 shows the results of the MT quality and translation speed among each method.

First, comparing the baseline with the statistical generations (Bottom-up and All Search), the MT quality of statistical generation improved in all evaluation metrics. Accordingly, the models of statistical MT are effective for improving the MT quality of example-based MT.

Next, comparing Bottom-up with All Search, the MT quality of bottom-up generation was slightly low. Bottom-up generation locally applies the translation model to a partial tree. In other words, the probability is calculated without word alignment linked to the outside of the tree. This result indicates that the results of bottom-up generation are not equal to the global optimal solution.

Comparing LM Only with the statistical generations, the MT quality of ranks A+B+C by subjective evaluation significantly decreased. This is because the n-gram language model used here does not consider output length, and shorter translations are preferred. Although the language model was effective to some degree, it could not evaluate the equivalence of the translation and the input sentence. Therefore, we concluded that the translation model is necessary for improving MT quality.

Finally, focusing on translation speed, the worst time for Bottom-up generation was dramatically faster than that for All Search. Bottom-up generation effectively uses shared nodes of the target tree, so it can improve translation speed. Therefore, bottom-up generation is suitable for tasks that require real-time processing, such as spoken dialogue translation.

5 Discussion

We incorporated example-based MT in models of statistical MT. However, some methods to obtain initial solutions of statistical MT by example-based MT have already been proposed. For example, Marcu (2001) proposed a method in which initial translations are constructed by combining bilingual phrases from translation memory, which is followed by modifying the translations by greedy decoding (Germann et al., 2001). Watanabe and Sumita (2003) proposed a decoding algorithm in which translations that are similar to the input sentence are retrieved from bilingual corpora and then modified by greedy decoding.

The difference between our method and these methods involves whether modification is applied. Our approach simply selects the best translation from candidates that are output from example-based MT. Even though example-based MT can output appropriate translations to some degree, our method assumes that the candidates contain a globally optimal solution. This means that the upper bound of MT quality is limited by the example-based transfer, so we have to improve this stage in order to further improve MT quality. For instance, example-based MT can be improved by applying an optimization algorithm that uses an automatic evaluation of MT quality (Imamura et al., 2003).

6 Conclusions

This paper demonstrated that example-based MT can be improved by incorporating it in models of statistical MT. The example-based MT used in this paper is based on syntactic transfer, so word reordering is achieved in the transfer module. Using this feature, the best translation was selected by using only a lexicon model and an n-gram language model. In addition, bottom-up generation achieved faster translation speed by using the tree structure of the target sentence.

Acknowledgements

Ruigo-shin-jiten.

References

Yasuhiro Akiba, Taro Watanabe, and Eiichiro Sumita. 2002. Using language and translation models to select the best among outputs from multiple MT systems. In *Proceedings of COLING-2002*, pages 8–14.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Philip Clarkson and Ronald Rosenfeld. 1997. Statistical language modeling using the CMU-Cambridge toolkit. In *Proceedings of EuroSpeech 97*, pages 2707–2710.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the HLT Conference*, San Diego, California.

Ulrich Germann, Michael Jahr, Kevin Knight, Daniel Marcu, and Kenji Yamada. 2001. Fast decoding and optimal decoding for machine translation. In *Proceedings of 39th Annual Meeting of the Association for Computational Linguistics*, pages 228–235.

Kenji Imamura, Eiichiro Sumita, and Yuji Matsumoto. 2003. Feedback cleaning of machine translation rules using automatic evaluation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003)*, pages 447–454.

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

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

Genichiro Kikui, Eiichiro Sumita, Toshiyuki Takezawa, and Seiichi Yamamoto. 2003. Creating corpora for speech-to-speech translation. In *Proceedings of EuroSpeech 2003*, pages 381–384.

Daniel Marcu. 2001. Towards a unified approach to memory- and statistical-based machine translation. In *Proceedings of 39th Annual Meeting of the Association for Computational Linguistics*, pages 386–393.

Makoto Nagao. 1984. A framework of mechanical translation between Japanese and English by analogy principle. In *Artificial and Human Intelligence*, pages 173–180, Amsterdam: North-Holland.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Susumu Ohno and Masato Hamanishi. 1984. *Ruigo-Shin-Jiten*. Kadokawa, Tokyo. in Japanese.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318.

Eiichiro Sumita and Hitoshi Iida. 1991. Experiments and prospects of example-based machine translation. In *Proceedings of the 29th ACL*, pages 185–192.

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

Stephan Vogel, Ying Zhang, Fei Huang, Alicia Tribble, Ashish Venugopal, Bing Zhao, and Alex Waibel. 2003. The CMU statistical machine translation system. In *Proceedings of the 9th Machine Translation Summit (MT Summit IX)*, pages 402–409.

Taro Watanabe and Eiichiro Sumita. 2003. Example-based decoding for statistical machine translation. In *Proceedings of Machine Translation Summit IX*, pages 410–417.