Word Selection for EBMT based on Monolingual Similarity and Translation Confidence

Eiji Aramaki[†‡], Sadao Kurohashi[†‡], Hideki Kashioka[‡] and Hideki Tanaka[‡]

[†] Graduate School of Information Science and Tech. University of Tokyo
Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
{aramaki, kuro}@kc.t.u-tokyo.ac.jp

[‡] ATR Spoken Language Translation Research Laboratories
2-2 Hikaridai, Seika, Soraku, Kyoto 619-0288, Japan
{hideki.kashioka, hideki.tanaka}@atr.co.jp

Abstract

We propose a method of constructing an example-based machine translation (EBMT) system that exploits a content-aligned bilingual corpus. First, the sentences and phrases in the corpus are aligned across the two languages, and the pairs with high translation confidence are selected and stored in the translation memory. Then, for a given input sentences, the system searches for fitting examples based on both the monolingual similarity and the translation confidence of the pair, and the obtained results are then combined to generate the translation. Our experiments on translation selection showed the accuracy of 85% demonstrating the basic feasibility of our approach.

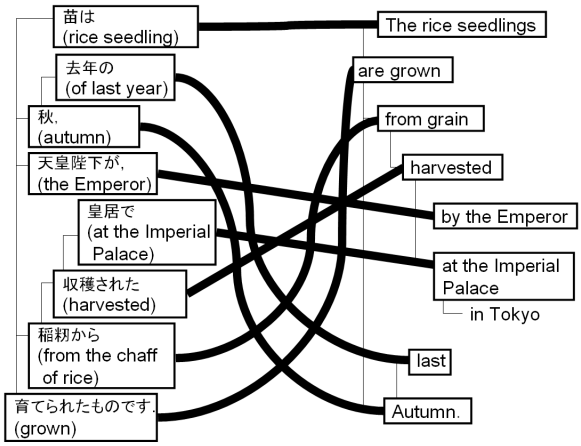

Figure 1: Translation Example (TE).

1 Introduction

The basic idea of example-based machine translation, or EBMT, is that translation examples similar to a part of an input sentence are retrieved and combined to produce a translation(Nagao, 1984). In order to make a practical MT system based on this approach, a large number of translation examples with structural correspondences are required. This naturally presupposes high-accuracy parsers and well-aligned large bilingual corpora.

Over the last decade, the accuracy of the parsers improved significantly. The availability of well-aligned bilingual corpora, however, has not increased despite our expectations. In reality, the number of bilingual corpora that share the same content, such as newspapers and broadcast news, has increased steadily. We call this type of corpus a content-aligned corpus. With these observations, we started a research project that covered all aspects of constructing EBMT systems starting from using a content-aligned corpus, i.e., a bilingual broadcast news corpus.

First, the sentences and phrases in the corpus are aligned across the two languages, and the pairs with high translation confidence are selected and stored in the translation memory. Then, translation examples are retrieved based on both the monolingual similarity and the translation confidence of the pair. Finally, these examples are combined to generate the translation.

This paper is organized as follows. The next section presents how to build the translation memory from a content-aligned corpus. Section 3 describes our EBMT system, paying special attention to the selection of translation examples. Section 4 reports experimental results of word selection, Section 5 describes related works, and Section 6 gives our conclusions.

石川県 (Ishikawa Prefecture) 輪島市で外国の大使や一般の参加者など千人あまりが急な斜面の棚田で田植えを体験する催しが行われました。輪島市白米町には (in Shiroyonemachi) 千枚田と呼ばれる大小 (of all various sizes) 二千百枚の棚田が急な斜面から海に向かって拡がっています。田植え体験は農作業を通して米作りの意義などを考えていこうという (thinking about the significance of the rice crop farming) 地球環境平和財団の呼び掛けで開かれたもので、海外三十四ヵ (34 oversea countries) 国の大使や書記官 (ambassador and secretary)、それに一般の参加者ら合わせておよそ千人が集まりました。田植えに使われた苗は去年の秋、天皇陛下が皇居で収穫された稲籾から育てたものです。参加者たちは裸足になって水田に足を踏み入れ地元に伝わる田植え歌に合わせて慣れない手つきで (unskillfully) 苗を植えていました。きょうの輪島市は雲が広がったもののまずまずの天気となり、出席された高円宮さまも海からの風に吹かれながら田植えに加わっていました。地球環境平和財団では今年の夏休みに全国の子どもたちを対象に草刈りや生きものの観察会を開く他、秋には稲刈体験を行なう予定にしています。(The weather in Wajima City was not bad. Prince Takamadonomiya joined the rice-planting feeling the wind from the sea. The private Foundation for Global Peace and Environment is planning to organize watching wildlife and mowing events in summer vacation and a harvesting event in autumn.)

Ambassadors and diplomats from 37 countries took part in a rice planting festival on Sunday in small paddies on steep hillsides in Wajima, central Japan. About one-thousand people gathered at the hill, where some two-thousand 100 miniature paddies, called Senmaida, stretch toward the Sea of Japan. The event was organized by the private Foundation for Global Peace and Environment. The rice seedlings are grown from grain harvested by the Emperor at the Imperial Palace in Tokyo last autumn. Barefoot participants waded into the paddies to plant the seedlings by hand while singing a local folk song about the practice of rice planting.

* Underlined phrases and sentences have no parallel expressions in the other language.

Figure 2: NHK News Corpus.

2 Building Translation Memory

In EBMT, an input sentence can hardly be translated by a single translation example, except when an input is extremely short or is a typical domain-dependent sentence. Therefore, two or more translation examples are used to translate parts of the input and are then combined to generate a whole translation. Syntactic information is useful for composing example fragments.

In this paper, we call a structurally aligned bilingual sentence pair a *translation example* or *TE* (Figure 1). This section presents our method for building TEs from a content-aligned corpus.

Since the bilingual corpus used in our project does not contain literal translations, automatic parsing and alignment inevitably contain errors. Therefore, we selected highly likely TEs to make a *translation memory*.

2.1 NHK News Corpus

We used a bilingual news corpus compiled by the NHK broadcasting service (NHK News Corpus), which consists of about 40,000 Japanese-English article pairs covering a five-year period. The average number of Japanese sentences in an article is 5.2, and that of English sentence is 7.4. Table 2 shows an example of an article pair.

As shown in Table 2, an English article is not a literal translation of a Japanese article, although their contents are almost parallel.

2.2 Sentence Alignment

We used a DP matching for bilingual sentence alignment, where we allow the matching of 1-to-1, 1-to-2, 1-to-3, 2-to-1 and 2-to-2 Japanese and English sentence pairs. This matching covered 84% of the following evaluation set. We selected 96 article pairs for the evaluation of sentence and phrase alignment, and we call this the evaluation set. We use the following score for matching, which is based on a ratio of corresponding content words (*WCR*: content Word Corresponding Ratio).

$$\mathrm{WCR} = \frac{W_d}{W_j + W_e}, \tag{1}$$

where W_j is the number of Japanese content words in a unit, W_e is the number of English content words, and W_d is the number of content words whose translation is also in the unit, which is found by translation dictionaries.

We used the EDR electronic dictionary, EDICT, ENAMDICT, the ANCHOR translation dictionary, and

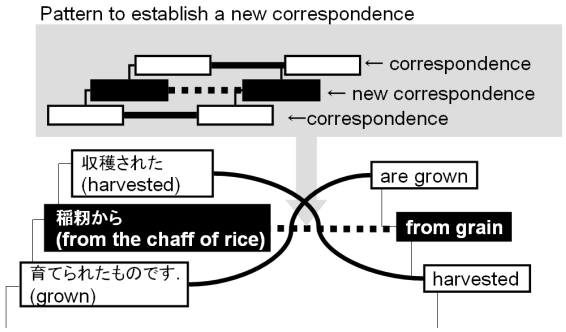

Figure 3: Handling of Remaining Phrases.

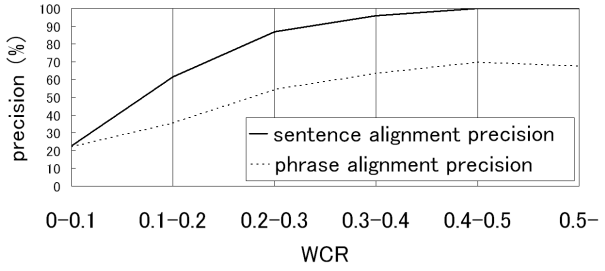

Figure 4: WCR and Precision.

the EIJIRO translation dictionary. These dictionaries have about two million entries in total.

On the evaluation data, the precision of the sentence alignment (defined as follows) was 60.7%.

$$\text{precision} = \frac{\text{\# of correct system outputs}}{\text{\# of system outputs}} \qquad (2)$$

Among types of a corresponding unit, the precision of 1-to-1 correspondence was the best, at 77.5%. Since a 1-to-1 correspondence is suitable for the following phrase alignment, we decided to use only the 1-to-1 correspondence results.

2.3 Phrase Alignment

The 1-to-1 sentence pairs obtained in the previous section are then aligned at phrase level by the method based on (Aramaki et al., 2001). The method consists of the following pre-process and two aligning steps.

Pre-process: Conversion to phrasal dependency structures.

First, the phrasal dependency structures of the sentence pair are estimated. The English parser returns a word-based phrase structure, which is merged into a phrase sequence by the following rules and converted into a dependency structure by lifting up head phrases.

Table 1: Number of TEs.

Corpus	*WCR*	# of TEs
NHK News	0.3–0.4	18290
	0.4–0.5	6975
	0.5–	2314
White Paper	—	2225
SENSEVAL	—	6920

1. Function words are grouped with the following content word.
2. Adjoining nouns are grouped into one phrase.
3. Auxiliary verbs are grouped with the following verb.

The Japanese parser outputs the phrasal dependency structure of an input, and that is used as is. We used The Japanese parser KNP (Kurohashi and Nagao, 1994) and The English nl-parser (Charniak, 2000).

Step 1: Estimation of basic phrasal correspondences.

We started with the word-level alignment to get the basic phrasal alignment. We used translation dictionaries for this process. The word sense ambiguity in the dictionaries is resolved with a heuristics that the most plausible correspondence is near other correspondences.

Step 2: Expansion of phrasal correspondences.

Finally, the remaining phrases, which were not handled in the step 1, are merged into a neighboring phrase correspondence or are used to establish a new correspondence, depending on the surrounding existing correspondences. Figure 3 shows an example of a new correspondence established by a structural pattern.

These procedures can detect the phrasal alignments in a pair of sentences as shown in Figure 1.

For phrase alignment evaluation, we selected all of the 145 sentence pairs that had 1-to-1 correspondences form the evaluation set and gave correct content word correspondences to these pairs. The phrase correspondences detected by the system were judged correct when the correspondences include the manually given content word correspondences.

Based on this criterion, the precision of phrase alignment was 50%. Then, we found a correlation between the phrase alignment precision and *WCR* of parallel sentences as shown in Figure 4. Furthermore, the precision of sentence alignment and *WCR* also have a correlation. Since their performances nearly reaches their limits when *WCR* is 0.3, we decided to use parallel sentences whose *WCR* is 0.3 or greater as TEs.

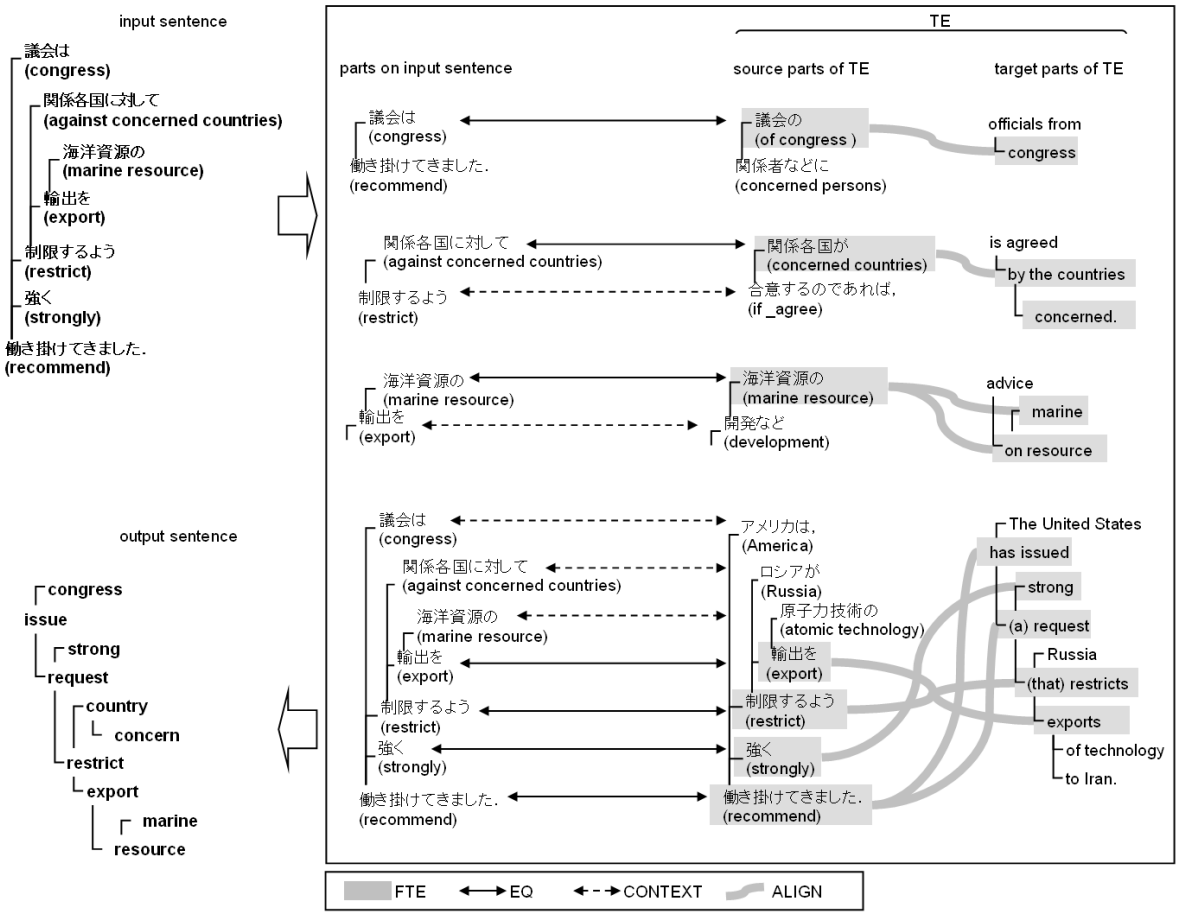

Figure 5: Example of Translation.

2.4 Building Translation Memory

As explained in the preceding sections, among sentence-aligned and phrase-aligned NHK News articles, TEs with a 1-to-1 sentence correspondence and whose *WCR* is 0.3 or greater are registered in the translation memory. Table 1 shows the number of TEs for each *WCR* range.

In addition, the Bilingual White Paper and Translation Memory of SENSEVAL2 (Kurohashi, 2001) were also phrase-aligned and registered in the translation memory. Sentence alignments are already given for these corpora. Since their parallelism are fairly high and the accuracies of their phrase alignments are more than 70%, we utilized all phrase-aligned sentence pairs as TEs (Table 1).

3 EBMT System

Our EBMT system translates a Japanese sentence into English. A Japanese input sentence is parsed and transformed into a phrase-based dependency structure. Then, for each phrase, an appropriate TE is retrieved from the translation memory that is most suitable for translating the phrase (and its neighboring phrases). Finally, the English expressions of the TEs are combined to produce the final English translation (Figure 5).

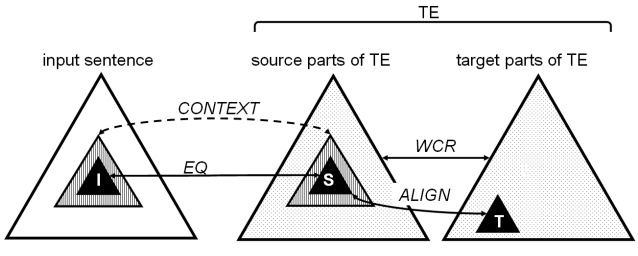

Figure 6: Selection of a TE.

This section describes our EBMT system, mainly the TE selection part.

3.1 Basic Idea of TE Selection

The basic idea of TE selection is shown in Figure 6. When a part of the input sentence and a part of the TE

source language sentence have an equal expression, the part of the input sentence is called *I* and the part of the TE source language sentence is called *S*. A part of the TE target language corresponding to *S* is called *T*. The pair *S* and *T* is called fragment of TE (FTE).

I, *S* and *T* have to meet the following conditions, as a natural consequence of the fact that *S*-*T* is used for translating *I*.

1. *I*, *S* and *T* are each structurally connected phrases.
2. *I* is equal to *S* except for function words at the boundaries.
3. *S* corresponds to *T* completely, that is, all phrases in *S* and *T* are aligned.

It might be the case that for an *I*, two or more FTEs that meet the above conditions exist in the translation memory. Our method takes into account the following relations among *I*-*S*-*T* to select the best FTE:

1. The largest pair of *I* and *S*.
2. The similarity between the surroundings of *I* and these of *S*.
3. The confidence of alignment between *S* and *T*.

The following sections concretely present how to calculate these criteria. For simplicity of explanation, we call a set of phrasal correspondences between *S* and *T*, *EQ*; that neighboring *EQ*, *CONTEXT*; that between *S* and *T*, *ALIGN* (Figure 6).

3.2 Monolingual Similarity between Japanese Expressions

The equality between *I* and *S* is a sum of the equality score of each phrase correspondence in *EQ*, which is calculated as follows:

$$EQUAL(i) = \frac{\sum S_{cont} \times 2}{\#_{cont}} + 0.2 \times \frac{\sum S_{func} \times 2}{\#_{func}}, \quad (3)$$

where $\#_{cont}$ is the number of content words in the phrase correspondence, $\#_{func}$ is the number of function words, S_{cont} is the equality between content words, and S_{func} is the equality between function words. S_{cont} and S_{func} are given in Table 2.[1]

Usually, the equality score between *I* and *S* is equal to the number of phrases in *I* (the number of phrase correspondences in *EQ*), but sometimes these are slightly different, depending on the conjugation type and function words.

[1]All constant values in Table 2 and formulas were decided based on preliminary experiments.

On the other hand, the similarity between the surroundings of *I* and those of *S* is a sum of the similarity score of each phrase correspondence in *CONTEXT*, which is calculated as follows:

$$SIM(i) = \left\{ \sum \frac{S_{cont} \times 2}{\#_{cont}} + 0.2 \times \frac{\sum S_{func} \times 2}{\#_{func}} \right\} \times S_{connect}. \quad (4)$$

Basically the calculation of *SIM* and *EQUAL* is the same, except that *SIM* considers the relation type between the phrase in *I* and its outer phrase by $S_{connect}$. When the relation is the same, the influence of the surrounding phrases must be large, so $S_{connect}$ is set to 1.0; when the relation is not the same, $S_{connect}$ is set to 0.5. The relations between phases are estimated by the function word or conjugation type of the dependent phrase.

The monolingual similarity between Japanese expressions *I* and *S* is calculated as follows:

$$\sum_{i \in EQ} EQUAL(i) + \sum_{i \in CONTEXT} SIM(i). \quad (5)$$

3.3 Translation Confidence of Japanese-to-English Alignment

The translation confidence of phrase alignment between *S* and *T* is the sum of the confidence score of each phrase correspondence in *ALIGN*, *CONF(i)* in Table 2, and it is weighted by the *WCR* of the parallel sentences.

As a final measure, the score of *I*-*S*-*T* is calculated as follows:

$$\left\{ \sum_{i \in EQ} EQUAL(i) + \sum_{i \in CONTEXT} SIM(i) \right\} \times \left\{ \sum_{i \in ALIGN} CONF(i) \right\} \times WCR. \quad (6)$$

3.4 Search Algorithm of FTE

For each phrase (*P*) in an input sentence, the most plausible FTE is retrieved by the following algorithm:

1. FTEs are retrieved from the translation memory, in which a Japanese phrase matches *P*, and it is aligned to an English phrase. (that is, these are FTEs that meet the basic conditions for translation in Section 3.1).
2. For each FTE obtained in the previous step, it is checked whether the surrounding phrase of *P* and that of FTE are the same or similar, phrase by phrase, and the largest *I*-*S*-*T* that meets the basic conditions is detected.

Table 2: Parameters for Similarity and Confidence Calculation.

S_{cont}	1.1	exact match
	1.0	stem match
	0.5 × S_{ntt} + 0.3	thesaurus match
	0.3	POS match
	0	otherwise
* S_{ntt} is a similarity calculated based on NTT thesaurus(Ikehara et al., 1997) (max = 1).		
S_{func}	1.1	exact match
	1.0	stem match
	0	otherwise
CONF(i)	1.0	all content words in alignment i correspond to each other in dic
	0.8	some content words in alignment i correspond to each other in dic
	0.5	otherwise

3. The score of each *I-S-T* is calculated, and the best *I-S-T* (*S-T* is the FTE) is selected as the FTE for *P*.

As a result of detecting FTEs for phrases in the input, two FTEs starting from the different phrase might overlap each other. In such a case, we employed a greedy search algorithm that adopts the higher score FTE one by one; therefore, each previously adopted FTE is only partly used for translation.

On the other hand, when no FTE is obtained for an input phrase, a translation dictionary is utilized (when the phrase contains two or more content words, the longest matching strategy is used for dictionary look-up). When two or more possible translations are given from the dictionary, the most frequent phrase/word in the NHK News Corpus is adopted.

Figure 5 shows examples of FTEs detected by our method.[2]

3.5 Generating a Target Sentence

The English expressions in the selected FTEs are combined, and the English dependency structure is constructed. The dependency relations in FTEs are preserved, and the relation between the two FTEs is estimated based on the relation of the input sentences. Figure 5 shows an example of a combined English dependency structure.

When a surface expression is generated from its dependency structure, its word order must be selected properly. This can be done by preserving the word order in FTEs and by ordering FTEs by a set of rules governing both the dependency relation and the word-order.

The module for controlling conjugation, determiner, and singular/plural is not yet implemented in our current MT system.

[2]As the bottom example in Figure 5 shows, EBMT can easily handle head-switching translation by using an FTE that contains all of the head-switching phenomena in it.

4 Experiments

For evaluation, we selected 50 sentence pairs from the NHK News Corpus that were not used for the translation memory. Their source (Japanese) sentences were translated by our EBMT system, and the selected FTEs were evaluated by hand, referring to the target (English) sentences.

A phrase by phrase evaluation was done to judge whether the English expression of the selected FTE was good or bad. The accuracy was 85.0%.

In order to investigate the effectiveness of each component of FTE selection, we compared the following four methods:

1. EQCONTEXTALIGN: The proposed method.

2. EQALIGN: FTE score is calculated as follows, without the *CONTEXT* similarity:

$$\sum_{i \in EQ} EQUAL(i) \times \sum_{i \in ALIGN} CONF(i) \times WCR. \quad (7)$$

3. EQCONTEXT: FTE score is calculated as follows, without the *ALIGN* confidence:

$$\sum_{i \in EQ} EQUAL(i) + \sum_{i \in CONTEXT} SIM(i). \quad (8)$$

4. DICONLY: Word selection is based only on dictionaries and frequency in the corpus.

The accuracy of each method is shown in Table 3, and the results indicate that the proposed method, EQCONTEXTALIGN, is the best, that is, using context similarity and align confidence works effectively. Figure 7

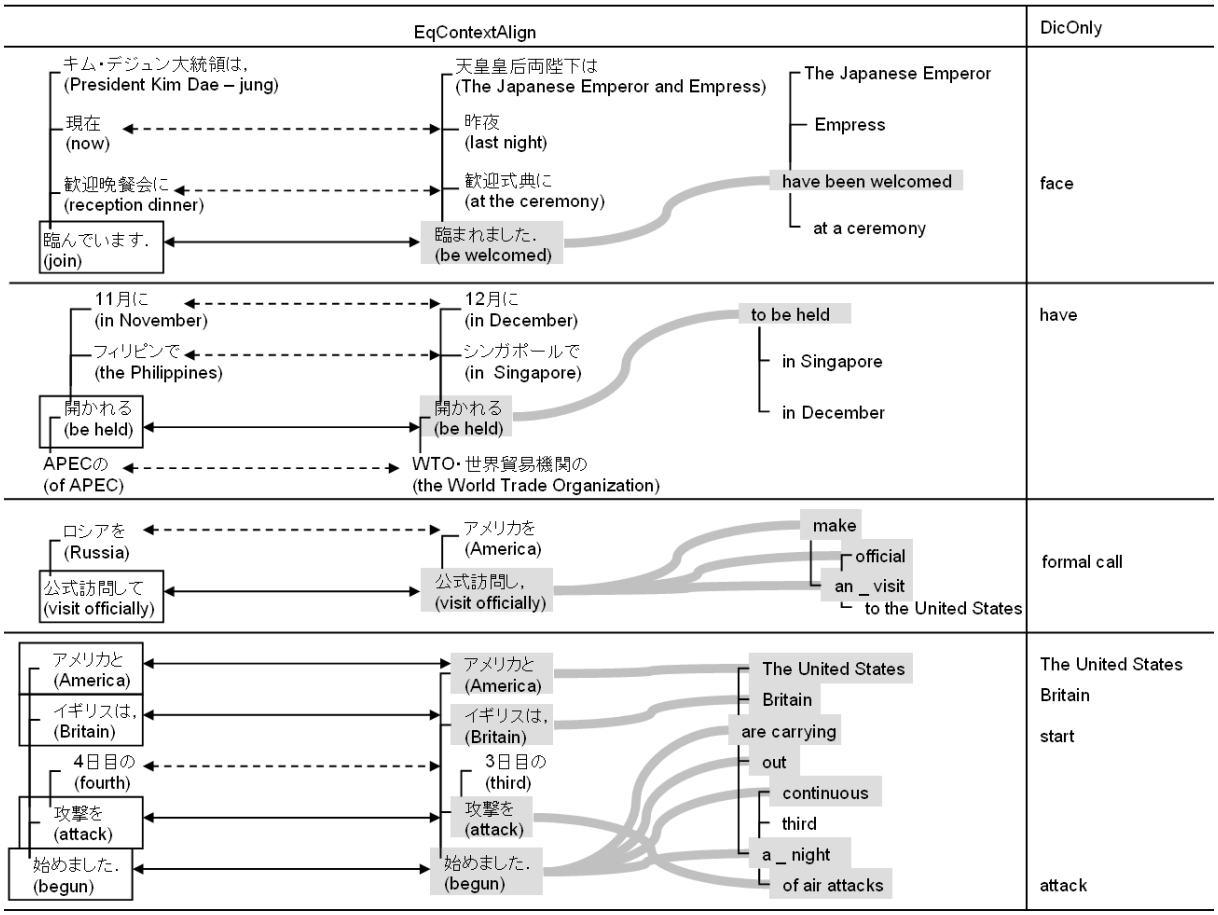

Figure 7: Word Selection by EQCONTEXTALIGN and DICONLY.

Table 3: Experimental Results.

	Good	Bad	Accuracy
EQCONTEXTALIGN	268	47	85.0%
	(246)	(35)	(87.5%)
EQALIGN	254	61	80.6 %
	(233)	(48)	(82.9%)
EQCONTEXT	234	80	74.2%
	(213)	(68)	(75.8%)
DICONLY	232	83	73.6%

* Values in brackets indicate the accuracy only for FTEs, excluding cases in which the dictionary was used as a backup.

shows examples of EQCONTEXTALIGN and DICONLY. EQCONTEXTALIGN usually selects appropriate words, compared to DICONLY.

When there are no plausible translation examples in the translation memory, the system selects a low-similarity or low-confidence FTE. However we believe this problem will be resolved as the number of translation examples increases, since the News Corpus is increasing day by day.

5 Related Work

The idea of example based machine translation systems was first proposed by (Nagao, 1984), and preliminary systems that appeared about ten years (Sato and Nagao, 1990; Sadler and Vendelmans, 1990; Maruyama and Watanabe, 1992; Furuse and Iida, 1994) showed the basic feasibility of the idea.

Recent studies have focused on the practical aspects of EBMT, and this technology has even been applied to some restricted domains. The work in (Richardson et al., 2001; Menezes and Richardson, 2001) addressed

the problem of technical manual translation in several languages, and the work of (Imamura, 2002) dealt with dialogues translation in the travel arrangement domain. These works select the translation example pairs based solely on the source language similarity. We believe this is partly due to the high parallelism found in their corpora.

Our work targets a more general corpus of wider coverage, i.e., the broadcast news collection. Generally available corpora like the one we use tend to be more freely translated and suffer from lower parallelism. This compelled us to use the criterion of translation confidence, together with the criterion of monolingual similarity used in the previous works. As we showed in this paper, this metric succeeded in meeting our expectations.

6 Conclusion

In this paper, we described operations of the entire EBMT process while using a content-aligned corpus, i.e., the NHK Broadcast Corpus. In this process, one of the key problems is how to select plausible translation examples. We proposed a new method to select translation examples based on source language similarity and translation confidence. In the word selection task, the performance is highly accurate.

Acknowledgements

This work was supported in part by the 21st Century COE program "Information Science and Technology Strategic Core" at University of Tokyo and by a contract with the Telecommunications Advancement Organization of Japan, entitled "A study of speech dialogue translation technology based on a large corpus".

References

Eiji Aramaki, Sadao Kurohashi, Satoshi Sato, and Hideo Watanabe. 2001. Finding translation correspondences from parallel parsed corpus for example-based translation. In *Proceedings of MT Summit VIII*, pages 27–32.

Eugene Charniak. 2000. A maximum-entropy-inspired parser. In *In Proceedings of NAACL 2000*, pages 132–139.

Osamu Furuse and Hitoshi Iida. 1994. Constituent boundary parsing for example-based machine translation. In *Proceedings of the 15th COLING*, pages 105–111.

Satoru Ikehara, Masahiro Miyazaki, Satoshi Shirai, Akio Yokoo, Hiromi Nakaiwa, Kentarou Ogura, and Yoshifumi Oyama Yoshihiko Hayashi, editors. 1997. *Japanese Lexicon*. Iwanami Publishing.

Kenji Imamura. 2002. Application of translation knowledgeacquired by hierarchical phrase alignment for pattern-based mt. In *Proceedings of TMI-2002*, pages 74–84.

Sadao Kurohashi and Makoto Nagao. 1994. A syntactic analysis method of long Japanese sentences based on the detection of conjunctive structures. *Computational Linguistics*, 20(4).

Sadao Kurohashi. 2001. Senseval2 Japanese translation task. In *Proceedings of SENSEVAL2*, pages 37–40.

Hiroshi Maruyama and Hideo Watanabe. 1992. The cover search algorithm for example-based translation. In *Proceedings of TMI-1992*, pages 173–184.

Arul Menezes and Stephen D. Richardson. 2001. A best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora. In *Proceedings of the ACL 2001 Workshop on Data-Driven Methods in Machine Translation*, pages 39–46.

Makoto Nagao. 1984. A framework of a mechanical translation between Japanese and english by analogy principle. In *In Artificial and Human Intelligence*, pages 173–180.

Stephen D. Richardson, William B. Dolan, Arul Menezes, and Monica Corston-Oliver. 2001. Overcoming the customization bottleneck using example-based mt. In *Proceedings of the ACL 2001 Workshop on Data-Driven Methods in Machine Translation*, pages 9–16.

V. Sadler and R. Vendelmans. 1990. Pilot implementation of a bilingual knowledge bank. In *Proeedings of the 13th COLING*, pages 449–451.

Satoshi Sato and Makoto Nagao. 1990. Toward memory-based translation. In *Proceedings of the 13th COLING*, pages 247–252.