Example-based Machine Translation without Saying Inferable Predicate

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

†Graduate School of Information Science and Tech. University of Tokyo
{aramaki, kuro}@kc.t.u-tokyo.ac.jp

‡ATR Spoken Language Translation Research Laboratories
hideki.kashioka@atr.jp

†††Science and Technical Research Laboratories of NHK
tanaka.h-ja@nhk.or.jp

Abstract

For natural translations, a human being does not express predicates that are inferable from the context in a target language. This paper proposes a method of machine translation which handles these predicates. First, to investigate how to translate them, we build a corpus in which predicate correspondences are annotated manually. Then, we observe the corpus, and find alignment patterns including these predicates. In our experimental results, the machine translation system using the patterns demonstrated the basic feasibility of our approach.

1 Introduction

With the rapid growth of the Internet, the availability of electronic texts is increasing day by day. View of this, much attention has been given to data-driven machine translation, such as example-based machine translation (Nagao, 1984) and statistical machine translation (Brown et al., 1993). However, previous studies have mainly focused on parallel translations.

In reality, however, a human being often does not make a perfectly parallel translation. In the following example, T_{human} is a human translation of an input sentence S, and T_{mt} is one of our machine translation system.

S: Canada-de hirakareta tsuusyou-kaigi-de ...
(a trade conference that was held in Canada)
T_{mt}: At a trade conference held in Canada ...
T_{human}: At a trade conference in Canada ...

A machine translation system tends to translate word for word, as shown in T_{mt}. On the other hand, a human being does not explicitly translate the underlined verb phrase (VP) "hirakareta (*be held*)" as shown in T_{human}. The reason why the underlined phrase is not expressed is that a human avoids redundant expressions, and prefers a compact translation without it. We call such a phrase which is not expressed in translation *null-align* phrase in this paper. Besides the fact verb phrases are sometimes null-aligned, the difficulty of VP-alignments has been pointed out (Aramaki et al., 2001).

For this reason, in order to investigate how to translate VPs, we built a VP-aligned-corpus with two types of information annotated: (1) for each phrase, whether the phrase is a VP or not, (2) for each VP, where the VP in one language corresponds in the other. In this paper, we analyze the VP-aligned-corpus and suggest a method for achieving appropriate VP translations.

Though the proposed method does not depend on language pairs and translation directions, this paper describes Japanese-English translation.

This paper is organized as follows. The next section presents how to build the VP-aligned-corpus. Section 3 reports several observations of the VP-aligned-corpus. Section 4 describes how to achieve appropriate VP translations. Then, Section 5 reports experimental results, Section 6 describes related works, and Section 7 presents

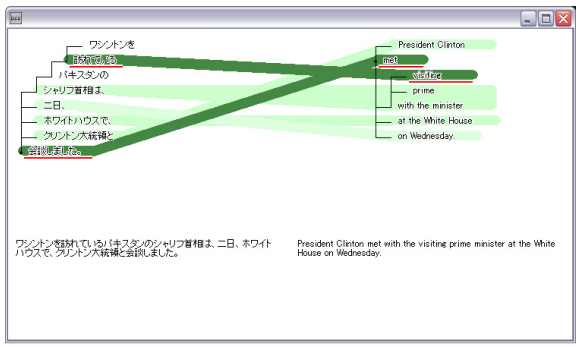


Figure 2: Annotation Tool

our conclusions.

2 VP-aligned-corpus

The VP-aligned-corpus is built using the following method: First, all of the sentence pairs in the corpus are automatically converted into phrasal dependency structures, and their phrasal alignments are estimated. Next, annotators modify the correspondences. In this section, we describe a corpus for the annotation and its annotation process.

2.1 NHK-news-corpus

To build a VP-aligned-corpus, we need a bilingual corpus consisting of natural translations. We used a bilingual news corpus compiled by the NHK broadcasting service (NHK News Corpus). It consists of about 40,000 Japanese articles (from a five-year period) and English ones which are translations of Japanese articles by humans. The average number of Japanese sentences in one article is 5.2, and that of English sentence is 7.4. Figure 1 shows an example of an article pair. In Figure 1, the underlined phrases and sentences have no parallel expressions in the other language. A large number of underlined expressions indicates that the Japanese articles are freely translated to be natural as English news.

2.2 Annotation Process

The annotation process consists of the following four steps:

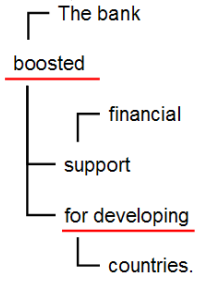
STEP 1: Estimation of Sentence Alignment
We use DP matching for bilingual sentence alignment based on a translation dictionary (two million entries in total). Next, we extract 1-to-1 sentence pairs. For the evaluation data (96 articles), the precision of the sentence alignment was 77.5% (Aramaki et al., 2003).

STEP 2: Conversion to Phrasal Dependency Structures
The phrasal dependency structures of the sentence pairs are estimated by parsers. The English parser (Charniak, 2000) returns a word-based phrase structure, which is merged into phrases by the following rules and converted into a dependency structure by deciding head phrases (Aramaki et al., 2003).

1. Function words are attached to their following content words.
2. Compound nouns are attached into one phrase.
3. Auxiliary verbs are attached to the main verb.

The Japanese parser KNP (Kurohashi and Nagao, 1994) outputs the phrasal dependency structure, and that is used as is.

STEP 3: Phrasal Annotation
VPs are annotated in the phrasal dependency structures. We define a VP as a phrase that contains (1) a verb or (2) an adjective that has an argument as its child. In the definition, we also regard a phrase which contains a gerund as a VP. For example, the following sentence has two VPs (“for developing” and ‘boosted”):



STEP 4: Correspondence Annotation: For each Japanese VP, annotators mark its corresponding phrases in the English sentence. As mentioned before, Japanese VPs do not

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.

Figure 1: NHK-news-corpus

always correspond to English VPs. We also allow annotators to mark the phrases that are not VPs (for example, a NP or a PP). In addition, when a Japanese VP has no parallel expressions, annotators mark **VP-**ϕ. In the same way, for each English VP, annotators make the annotations.

We annotated 5,500 sentence pairs. The annotation work was carried out using a GUI tool that can be operated by a mouse. Figure 2 shows the annotation tool and an annotated sentence pair. In this paper, we illustrate a sentence structure by locating its root node at the left as shown in Figure 2.

3 Analysis of a VP-aligned-corpus

In a VP-aligned-corpus, Japanese VPs do not always correspond to English VPs literally. We classify and count annotated correspondences from the view point of where Japanese phrases correspond in English (Table 1). As shown in Table 1, since the ratio of correspondences that are not VP-VP is more than 40%, so we cannot consider them as exceptional phenomena.

This section describes the classification.

Table 1: Classification of VP-correspondences

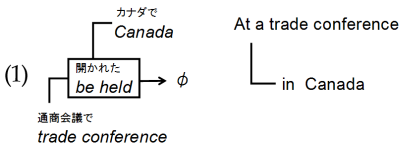
Classification (Japanese: English)	#
VP-VP	9779
VP-ϕ	6831
VP-PP, VP-NP	*710*
OTHERS	*316*

* The numbers in *Italic* are estimated automatically because the annotated information tells whether phrases are VPs or not.

1. **VP-VP**: A Japanese-VP corresponds to an English-VP.
   We paid little attention to **VP-VP**s in this paper, because these correspondences are estimated by conventional alignment methods.

2. **VP-**ϕ: A Japanese-VP has no parallel expressions in the phrase level.
   VP-ϕs arise for the two reasons: (1) the sentence alignments failed and the Japanese-VP has a parallel expression in another English sentence, or (2) the Japanese-VP occurs in a context that allow it to be null-aligned.

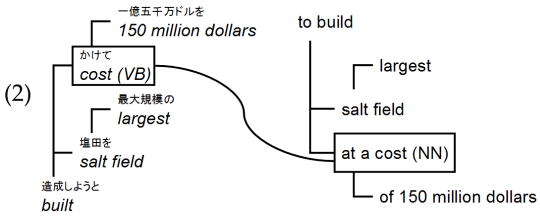
   The latter example, has been already shown in Section 1, is illustrated as follows:



(At a trade conference in Canada ...[1])

As I mentioned before, "hirakareta (*be held*)" in the above context is redundant to translate. We present a more detailed classification of this type in Section 4.

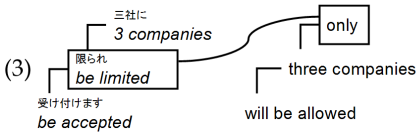
3. **VP-PP, VP-NP**: A Japanese-VP corresponds to an English-NP or PP.
   A Japanese VP is sometimes translated into a "NP" or a "preposition + NP." In the following example, "kakete (*cost or spend*)" is translated into a PP.



(to build the largest salt field
at a cost of 150 million dollars)

The following discussion does not deal with these cases, because they are also estimated by conventional alignment methods.

4. **Others**: A Japanese-VP corresponds to a phrase in another category.
   In the following example, a Japanese VP is translated into an adverb .



(only three companies will be allowed)

We paid little attention to this type, because the number of this type is small as shown in Table 1.

Table 2: Judgements of CAPs

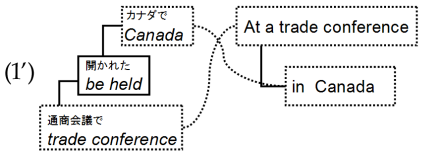
Judgement	Classification	#
good	P-CONTEXT	21
	C-CONTEXT	16
	BOTH-CONTEXT	19
	sum	56
bad	Parse error	3
	Alignment error	11
	Phrase chunking error	1
	Others	9
	sum	24

* The classifications in the good judgment (P, C, BOTH-CONTEXT) are mentioned in the next page.

4 Learning Null-aligned Translations

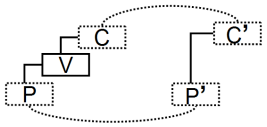
we concentrate **VP-**ϕ in this paper because the amount of it is the highest the other classifications except for **VP-VP**.

The observation of **VP-**ϕ leads to the fact that the surroundings of a **VP-**ϕ are parallel with each other. For example, (1) in the last section is aligned as follows:



We call such an alignment pattern consisting of three Japanese phrases and two English phrases a **Condensed Alignment Pattern** or simply a **CAP** in this paper[2].

The following is an image of a CAP.



If a null-aligned phrase in a CAP is always inferable and redundant to translate, we can regard CAPs as translation examples, and achieve compact translations using them.

In order to examine the above assumption, we randomly extracted 80 CAPs from the corpus. Then, we manually checked whether null-aligned phrases in CAPs are inferable (good) or not (bad) (Table 2). As a result, except for some errors

[1]The expressions in the branket are a part of an English sentence.

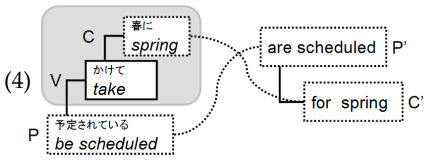
[2]There are reverse CAPs consisting of three Japanese phrase and two English phrases in the corpus. However we paid no attention to them because this paper deals with translation in the Japanese-English direction

(alignment errors, parse errors and so on) almost all CAPs are appropriate as translation examples. Therefore, we can use them as translation examples.

However, if we regard an entire CAP as a translation example, it can be used only in the case where it is equal to the input sentence. To cope with this problem, we estimate unnecessary parts of CAPs, and generalize them.

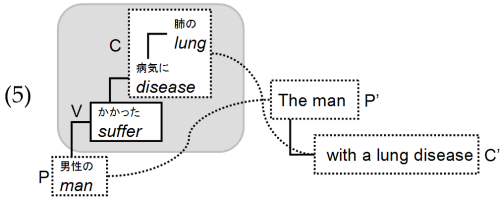
First, we classify the CAPs depending on whether its parent (**P**) or child (**C**) is the unnecessary context.

1. **C-CONTEXT**: only **C** is a necessary context.
   There is a case in which a null-aligned VP tied to its child (**C**) and its parent (**P**) is not a necessary context. In the following example, the Japanese **V** (*take*) performs as only a case-marker for **C** (*spring*).



(...are scheduled for spring...)

   In the above example, both a **V** and a **P** are VPs. On the other hand, in the following example, the **P** is a NP.



(The man with a lung disease ...)

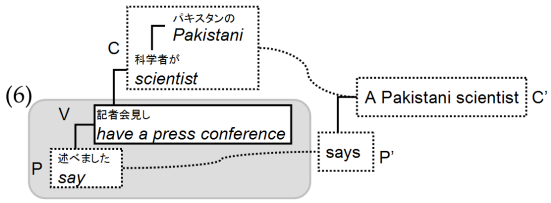
   In this example, a **V** is a null-aligned phrase, because it is associated with **C** (*disease*).

   Accordingly, we can see different linguistic phenomena depending on whether the **P** is a VP or not. However, we deal with them uniformly in from the structural view point.

2. **P-CONTEXT**: only **P** is a necessary context.

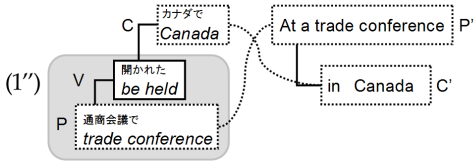
   In contrast, the following examples are cases in which a **C** and a **P** are tied each other. In this type, we can also see different linguistic phenomena depending on whether the **P** is a VP or not.

   When the **P** is a VP, the **P** and the **V** have similar meanings, because the **P** has the child phrase (**C**) instead of the **V**. In the following, **V** "kisya-kaiken-shi (*have a press conference*)" is a null-aligned phrase, then **P**"say" has **C**"a Pakistani scientist" as its child.



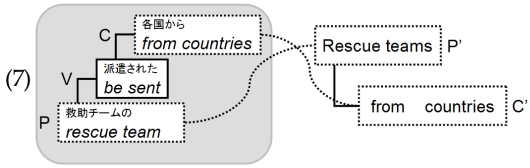
(A Pakistani scientist says ...)

   On the other hand, if the **P** is not a VP, the **P** associates with a **V**, and the **V** is a null-aligned phrase. (1') is an example of this type, and is illustrated as follows:



(At a trade conference in Canada ...)

3. **BOTH-CONTEXT**: Both a **P** and a **C** are necessary contexts.
   There is cases in which both a **P** and a **C** are necessary contexts. In such a case, we regard the entire CAP as a translation example. In the following example, both **C** (*each country*) and **P** (*rescue teams*) associate with **V**(*be sent*).



(Rescue teams from countries ...)

Table 3: Examples of CAP Fragments and their Frequencies

CAP Fragments: various Ps and V"*hiraku*(be held)"				
V	**P**	**P'**	Frequency	
hiraku	*kaigi*	conference	17	$=freq(P)$
	kaigou	meeting	2	
	syuukai	gathering	1	
CAP Fragments: various Cs and V"*hiraku*(be held)"				
C	**V**	**C'**	Frequency	
kaigi		meeting	5	
Canada		Canada	4	$=freq(C)$
Lyon	*hiraku*	Lyon	4	
Singapore		Singapore	3	
London		London	3	

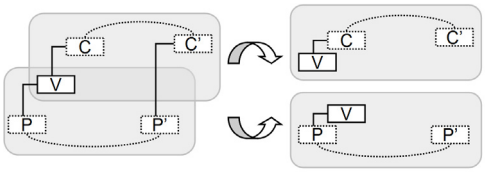


Figure 3: Fragments of a CAP

4.1 Estimation CAP Context

The classification in the above section is based on subjective judgments and estimations of all the CAP's contexts is difficult and insignificant. However, in the case that a **P** is obviously the context, we can find many CAPs that include **P**, **P'** and **V**. Therefore, we divide a CAP into two CAP-fragments and count their occurrences.

For example, Table 3 shows CAP fragments that include "hirakareta (*be held*)" from the Section 1 example. CAP fragments including "kaigi(*conference*)" occur 17 times, and those including "Canada (*Canada*)" occur only 4 times. Therefore, We can decide that "kaigi(*conference*)" is the context that allows "hirakareta (*be held*)" to be a null-aligned phrase.

The algorithm of the context estimation is follows:

1. For a phrase in a CAP, we decide the headword. For a NP, we regard the last noun as the headword. For a VP, we regard the main verb as the headword. Otherwise, we regard the entire phrase as the headword.

2. First, we divide CAP into two fragments (Figure 3). Then, we count their frequencies. Where, the frequency of (**P**,**P'**,**V**) is $freq(P)$, and that of (**C**,**C'**,**V**) is $freq(C)$.

3. After counting, for each CAP, if ($freq(P) > freq(C) \times 2$), a **P** is the context (P-CONTEXT)[3].

4. On the contrary, if ($freq(C) > freq(P) \times 2$), a **C** is the context (C-CONTEXT).

5. Otherwise, both a **P** and a **C** are contexts (BOTH-CONTEXT).

Table 5: Estimated CAP Context

	# of CAPs
P-CONTEXT	1120
C-CONTEXT	297
BOTH-CONTEXT	2802

5 Experiments

We evaluated our method from the following two view points: (1)how many CAPs were extracted (2)how much CAPs improved the translation accuracy.

5.1 CAP Extraction

We examined how many CAPs were extracted from our translation examples. The translation examples consist of 52,749 automatically aligned sentence pairs that were extracted from NHK News Corpus. As a result, 4,219 CAPs were extracted in total. It shows that we can extract one CAP from each 12 sentence pairs.

Table 5 shows the ratio of their estimated contexts. Most of them were estimated as BOTH-CONTEXT. However, 2,272 out of 2,802 BOTH-CONTEXTs were unique, then their $freq(C)$ and $freq(P)$ was 1. Therefore, if we got more translation examles, some of them would be reclassified into P or C-CONTEXT.

Table 4 shows examples of CAPs. As mentioned before, the judgments of the CAP context is too subjective to argue. We evaluated the propriety of context estimation in full translation tasks.

Table 4: Examples of CAP Contexts Estimation

C	V	P	C'	P'	freq(C):freq(P)
P-CONTEXT					
gakusei-ga *(students)*	kuwawattean *(join)*	okonaware-mashita *(held)*	students	held	1:2
kaityou-to *(President)*	kaidan-shi *(talk)*	itti-shimashita *(agree)*	President	agreed	1:5
gozen-no *(morning)*	kisyakaiken-de *(press conference)*	shimeshi-mashita *(told)*	in morning	told	1:10
C-CONTEXT					
kyouryoku-shi *(cooperate)*	susumeru-koto-ga *(plan)*	hituyou-dato *(need)*	to cooperate	need	4:2
kaigou-wo *(meeting)*	hiraku *(held)*	basyo-ni-tuite *(location)*	of meetings	the location	5:2
atumaru *(gather)*	mitooshi-de *(future)*	yosou-sarte-imasu *(be expected)*	to gather	are ecpected	3:1
BOTH-CONTEXT					
3-nishi-kan-to-suru *(3 days)*	houkou-de *(plan)*	susumete-imashita *(prepare)*	for a three day visit	was preparing	1:1
taishi-wo *(ambassador)*	yon-de *(call)*	motome-mashita *(ask)*	ambassador	asked	2:2
souri-daijin-ha *(Mr. Hashimoto)*	kisyakaiken-shi *(press conference)*	kyoutyou-shimashita *(said)*	Mr. Hashimoto	said	6:7

Table 6: BLEU Score

	Testset [240]	Subset [104]	Subset [14]
$BASELINE$	24.6	24.7	26.3
$CAPMT$	24.8 (+0.8%)	-	29.0 (+10.2%)
$CAPMT+$	25.0 (+1.6%)	25.7 (+4.0%)	-

* The numbers in brackets are the compared ratio to $BASELINE$, and the numbers in square are # of sentences.

5.2 Full Translation

We evaluated the CAP's improvements in full translation tasks using our Japanese-English translation system (Aramaki et al., 2003). The system produces a translation using translation examples which are the similar to the input sentence.

We compared the following three conditions.

1. $BASELINE$: the EBMT system (Aramaki et al., 2003) without CAP translation examples.

2. $CAPMT$: the EBMT system using both $BASELINE$'s translation examples and CAP translation examples which are not estimated their contexts.

3. $CAPMT+$: the EBMT system using both $BASELINE$'s translation examples and CAP translation examples which are estimated their contexts.

We evaluated them using BLEU(Papineni et al., 2002). BLEU is a score computes the ratio of N-gram for translation results found in reference translations. We used N=3.

We prepared a testset consisting of 240 lead (top) sentences randomly extracted from the NHK-news-corpus, and four references that were made by NHK's professional translators. Some sentences in the testset were translated without CAP translation examples. In such sentences outputs of $CAPMT$ and $CAPMT+$ are equal to ones of $BASELINE$. Therefore, we also compared translation results in the subsets consisting of translations which are different from $BASELINE$.

The results shown in Table 6, and Table 7 presents some of the results. Although the score of $CAPMT+$ was almost equal to one of $BASELINE$ in the entire testset, it had 4.0% improvement in the subset. In addition, its coverage was high, because its subset consisted of 104 sentences.

In contrast, the $CAPMT$'s subset consisted of only 14 sentences. Therefore, its improvement

[3]This threshold was determined by a preliminary experiment not to deteriorate the accuracy of the system.

Table 7: Translation Examples

REF	... quake struck areas along northeastern Afghanistan ...
BASELINE	... disaster area of the earthquake occurred in afghanistan northeast ...
CAPMT+	... disaster area of the earthquake in afghanistan northeast ...
REF	An air show in the US state of Maryland on the 14th ...
BASELINE	Air show was held in maryland of the united states on the 14th ...
CAPMT+	Air show in maryland of the united states on the 14th ...
REF	... summit due to be held on the 25th.
BASELINE	... summit meeting conducted on 25th.
CAPMT+	... summit meeting on 25th.

had little statistical significance.

6 Related Work

The field of data-driven machine translation concentrates mainly on the study of statistical machine translation (SMT) and example-based machine translation (EBMT).

In SMT, a system has to deal with a freely translated corpus in order to estimate the condensed alignment patterns (CAPs) in this paper. Since the corpus in this paper (NHK News Corpus) has a high perplexity (more than 100 in the IBM Model 4), it is a difficult task for the SMT system.

In EBMT, previous researches deal with restricted domains that have fairly high parallelism, such as, software manuals (Menezes and Richardson, 2001), business documents (Sato and Saito, 2002), White Papers (Watanabe et al., 2000) and so on. In such corpora, the condensed alignment patterns are rare.

7 Conclusion

In this paper, we describe the classification of verb phrase translations, and proposed the method to translate null-aligned verb phrases. On the theoretical side, the proposed method works well, because null-aligned translation examples improved translations in the half testset, as shown in the experimental results. However, for all of the testset, our method did not achieve extensive improvement. One of the reasons is that the amount of condensed alignment patterns is not enough. However we believe this problem will be resolved as the size of the corpus increases, because the News Corpus is increasing day by day.

Acknowledgements

This work was supported in part by the 21st Century COE program “Information Science and Technology Strategic Core” at University of Tokyo and also in part 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.

Eiji Aramaki, Sadao Kurohashi, Hideki Kashioka, and Hideki Tanaka. 2003. Word selection for ebmt based on monolingual similarity and translation confidence. In *Proceedings of the HLT-NAACL 2003 Workshop on Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, pages 57–64.

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

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

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).

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 *Artificial and Human Intelligence*, pages 173–180.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL 2002*, pages 311–318.

Kengo Sato and Hiroaki Saito. 2002. Extracting word sequence correspondences with support vector machine. In *Proceedings of the 19th COLING*, pages 870–876.

Hideo Watanabe, Sadao Kurohashi, and Eiji Aramaki. 2000. Finding structural correspondences from bilingual parsed corpus for corpus-based translation. In *Proceedings of the 18th COLING*, pages 906–912.