Bilingual Knowledge Extraction Using Chunk Alignment

Young-Sook Hwang
ATR SLT Research Labs
2-2-2 Hikaridai Seika-cho
Soraku-gun Kyoto
619-0288, JAPAN
youngsook.hwang@atr.jp

Kyonghee Paik
ATR SLT Research Labs
2-2-2 Hikaridai Seika-cho
Soraku-gun Kyoto
619-0288, JAPAN
kyonghee.paik@atr.jp

Yutaka Sasaki
ATR SLT Research Labs
2-2-2 Hikaridai Seika-cho
Soraku-gun Kyoto
619-0288, JAPAN
yutaka.sasaki@atr.jp

Abstract

In this paper, we propose a new method for effectively acquiring bilingual knowledge by exploiting the dependency relations among the aligned chunks and words. We use a monolingual dependency parser to automatically obtain dependency parses of target language using chunk and word alignment. For reducing the computational complexity of structural alignment, we use a bilingual dictionary and adopt a divide-and-conquer strategy. By sharing the dependency relations of a given source sentence, we automatically obtain a dependency parse of a target sentence that is structurally consistent with the source sentence. Moreover, we extract bilingual knowledge bases from translation correspondences of singletons to surface verb subcategorization patterns by exploiting the bilingual dependency relations. To acquire reliable ones, we take a stepwise filtering method based on statistical test.

1 Introduction

Corpus-based approaches based on bilingual corpora are promising for automatically acquiring translation knowledge (Alahawi et al,2000) (Brown et al,1993)(Imamura,2002)(Menezes et al,2001) (Utsuro et al,1993)(Watanabe et al,2000)(Yamada et al,2001). Most of statistical translation models based on IBM models(Brown et al,1993) are built from bilingual corpora without considering the structural aspects of the language(Brown et al,1993). They often output ungrammatical or unnatural translations.

(Yamada et al,2001) modeled the translation process from a parse tree of source language into a target language sentence. But the computational complexity during alignment is very high because it must handle hierarchical structures. Some methods use parsed sentences in parallel sentence-aligned corpora to extract transfer rules or examples (Aramaki et al,2001) (Watanabe et al,2000) (Menezes et al,2001) (Imamura,2002). However, parse-to-parse matching, which regards parsing and alignment as separate and successive procedures, suffers from grammatical inconsistency between languages. Moreover, it costs a lot to develop parsers of both the source and target language.

In this study, we propose a new method for effectively extracting translation knowledge which will reduce the computational complexity of alignment without losing the structural properties of each language. The main difference of our approach lies in exploiting a pair of a dependency parsed sentence and a POS tagged sentence as an input, and aligning those sentences in a chunk level as well as in a word level. By sharing the dependency relations of a source sentence, it is possible to automatically obtain a dependency parsed target sentence that will be structurally consistent with a given source sentence without using a parser of the target language. Ultimately, we can effectively acquire invaluable bilingual knowledge by exploiting the dependency relations among the aligned chunks and words.

The proposed method was implemented into the system that consists of two sub-systems, a bilingual alignment system and a knowledge extraction system. For aligning chunks and words, we utilize bilingual dictionaries and take a divide-and-conquer strategy. That is, after aligning chunks, the alignment of word level is consecutively tried in chunk pairs to reduce the computational complexity and improve alignment accuracy.

For extracting linguistic knowledge, we first obtain bilingual dependency parses by sharing the dependency relations between the chunk and word aligned sentences. Then we recursively traverse the dependency parses to get various bilingual knowledge from translation correspondences of singletons to surface verb sub-categorization patterns and apply a stepwise filtering method to obtain reliable ones.

As a case study, we apply the proposed method to Japanese and Korean. For evaluating the extracted knowledge, we measure the coverage and the average ambiguities of them on the Basic Travel Expression Corpus(Takezawa et al,2002).

From now on, we will show the details of bilingual chunk alignment and word alignment in section 2.1., knowledge extraction in section 2.2. and some experiments of extracting bilingual knowledge and evaluating them in section 3. Then, we will discuss some issues in bilingual knowledge extraction, and conclude this study.

2 J/K Bilingual Knowledge Extraction Using Chunk Alignment

As a case study, we present a chunk and a word alignment between Japanese and Korean sentences as well as a bilingual knowledge extraction method from the obtained alignments. Figure 1 illustrates the overview of the proposed method. The system consists of two modules: a bilingual chunk aligner and a knowledge extractor which are incorporated to obtain stable chunk and word alignment as well as reliable bilingual knowledge. In other words, using feedback from reliable bilingual knowledge, we obtain more accurate chunk and word alignment. Moreover, we reconstruct more reliable and rich bilingual dictionaries using the feedback from accurate alignment. The alignment and the knowledge extraction proceeds until there is no change in the resulting alignment.

2.1 Chunk alignment and Word Alignment

A chunk alignment module consists of 3 steps. Step 1: word alignment based on bilingual dictionaries; Step 2: statistical chunk alignment constrained by the previous word alignment results and chunk boundaries of the source language; and Step 3: statistical word alignment in the aligned chunk pairs. An example is shown in Figure 2.

As an input, our chunk and word aligner takes a pair of dependency parsed Japanese sentence and a POS tagged Korean sentence. First, it tries to make possible word alignment between the two languages by referring to a J/K dictionary [1](Figure 2 (a)). Consulting a J/K dictionary, we first find translations of the inflected forms of a given word. If the translation fails, then we try to find translations of the morphological bases. We also look for matches between components of multi-word expressions and individual words. To disambiguate the words that have more than one lexical candidate, we take some linguistic constraints: 1) The position of the corresponding word is within a given window size because the word order of each language is the same, 2) A functional word is translated into a corresponding functional word.

Second, during chunk alignment, we utilize structural cues as well as the initial lexical correspondences established in the previous step. Structural cues include source chunk structure and similarities between the two languages (i.e. use of case markers and SOV order). However, simple structural cues is limited to specific sentence pairs having similar structure because a sentence can be output with various expressions and structures. To solve this problem, we adopt a statistical chunk alignment model.

Basically, a source chunk is constrained by the dependency relations of a given source sentence and lexical correspondences and structural cues in a sentence pair also act as constraints. A chunk is represented by a combination of its head and tail morphemes and a sequence of POS tags in order to capture the morpho-syntactic characteristics of a chunk.

[1] J/K dictionary consists of 101,590 entries. All are tagged by part-of-speech at the morpheme level. The morphological base and inflectional form of each word are maintained.

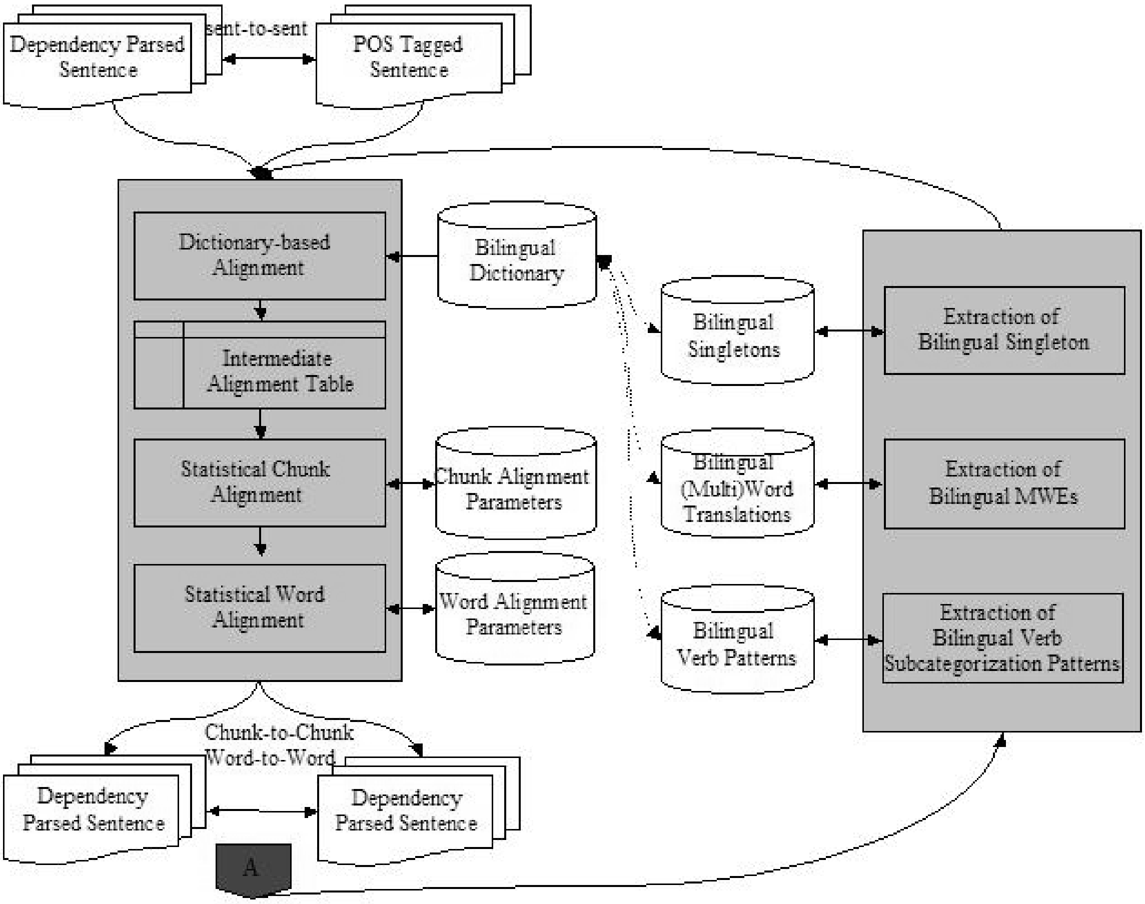

sent-to-sent
Dependency Parsed Sentence
POS Tagged Sentence
Dictionary-based Alignment
Intermediate Alignment Table
Statistical Chunk Alignment
Statistical Word Alignment
Bilingual Dictionary
Chunk Alignment Parameters
Word Alignment Parameters
Bilingual Singletons
Bilingual (Multi)Word Translations
Bilingual Verb Patterns
Extraction of Bilingual Singleton
Extraction of Bilingual MWEs
Extraction of Bilingual Verb Subcategorization Patterns
Chunk-to-Chunk
Word-to-Word
Dependency Parsed Sentence
Dependency Parsed Sentence
A

Figure 1: Bilingual Knowledge Acquisition via Chunk Alignment

Given a pair of Japanese and Korean sentence, $< J_c, K_w >$:

$$J_c = [w_J^{1..i}], \cdots, [w_J^{j..m}] = c_J^1 ... c_J^{m'} \qquad s.t. \quad 1 \le i \le j \le m, 1 \le m' \le m$$

$$K_w = w_K^1, \cdots, w_K^n \qquad s.t. \quad 1 \le n$$

where m and n denote the length of Japanese and Korean sentence respectively, $\overset{'}{m}$ is the number of Japanese chunks in J_c, and c_J^i denotes the type of the i^{th} Japanese chunk.

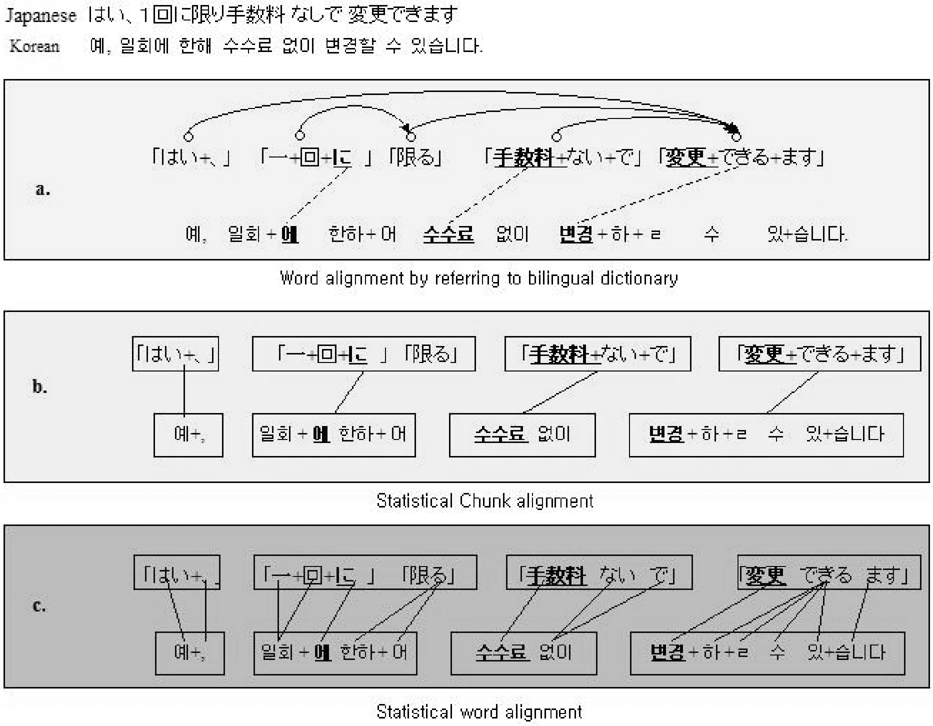
The chunk alignment model is structured as follows:

$$\begin{aligned} Pr(J_c, A_c|K_c) = & \quad Pr(J_c|A_c, K_c) \times Pr(A_c|K_c) \qquad (1) \\ = & \quad \prod_j Pr(c_J^j|c_J^{1..j-1}, a_1^j, K_c) \cdot Pr(a_j|c_J^{1..j-1}, a_1^{j-1}, K_c) \qquad (2) \end{aligned}$$

where K_c is a sequence of Korean chunks represented by $K_c = [w_K^{1..i'}], \cdots [w_K^{j'..l}] = c_K^1 ... c_K^{n'}$ under the condition $m' = n'$, and A_c denotes a sequence of alignment pairs (j, a_j).

In the model, we have two different probabilities: a chunk alignment probability $Pr(a_j|c_J^{1..j-1}, a_1^{j-1}, K_c)$ and a probability of chunk type correspondence $Pr(c_J^j|c_J^{1..j-1}, a_1^j, K_c)$. In the hidden Markov model, we assume a first-order dependence for the alignment a_j and that the probability of chunk type correspondence depends on the chunk at position a_j and the previous chunk c_J^{j-1}:

$$Pr(a_j|c_J^{1..j-1}, a_1^{j-1}, K_c) = Pr(a_j|a_{j-1}, c_J^{j-1}, c_K^{a_j}) \qquad (3)$$


Japanese はい、1回に限り手数料なしで 変更できます
Korean 예, 일회에 한해 수수료 없이 변경할 수 있습니다.
a.
「はい+、」「一+回+に 」「限る」「手数料+ない+で」「変更+できる+ます」
예, 일회+에 한하+여 수수료 없이 변경+하+ㄹ 수 있+습니다.
b.
「はい+、」「一+回+に 」「限る」「手数料+ない+で」「変更+できる+ます」
예+, 일회+에 한하+여 수수료 없이 변경+하+ㄹ 수 있+습니다
c.
「はい+、」「一+回+に 」「限る」「手数料 ない で」「変更 できる ます」
예+, 일회+에 한하+여 수수료 없이 변경+하+ㄹ 수 있+습니다

Word alignment by referring to bilingual dictionary

Statistical Chunk alignment

Statistical word alignment

Figure 2: An example of chunk and word alignment between Japanese and Korean

$$Pr(c_J^j|c_J^{1..j-1}, a_1^j, K_c) = Pr(c_J^j|c_J^{j-1}, a_j, c_K^{a_j}) \tag{4}$$

Putting ever thing together, we obtain the following HMM-based decomposition of $Pr(J_c|K_c)$:

$$Pr(J_c|K_c) = \sum_{A_c} \prod_j [Pr(a_j|a_{j-1}, c_J^{j-1}, c_K^{a_j}) \cdot Pr(c_J^j|c_J^{j-1}, a_j, c_K^{a_j})] \tag{5}$$

Parameters for alignment are estimated based on co-occurrences through unsupervised learning based on EM algorithm.

Word alignment is constrained by the aligned chunks (Figure 2 (c)). The statistical model for word alignment is similar to IBM model 2 (Brown et al,1993) except that a chunk pair is given as an input instead of a sentence pair. By constraining possible chunk alignment with partial lexical correspondences, structural cues, and taking a divide-and-conquer strategy, we can dramatically reduce the search space and computational complexity.

2.2 Extraction of Bilingual Knowledge

Before extracting knowledge, we first obtain bilingual dependency parses. They can be automatically obtained by using chunk aligned results and given dependency parses of source sentences (Figure 3 (a)). We assume that the dependencies among chunks are closely related to semantic relationship. If two chunks in the source sentence are dependent, the semantically corresponding chunks in the target sentence are also dependent on each other. Therefore, we can share the dependency relations between two languages even if we have dependency parses in one language.

Figure 3 shows an example of bilingual knowledge extraction by sharing the dependency relations of a given source sentence. We extract bilingual translation knowledge by exploiting the bilingual dependency parse trees. For extracting verb sub-categorization patterns, we first pivot the predicate of each

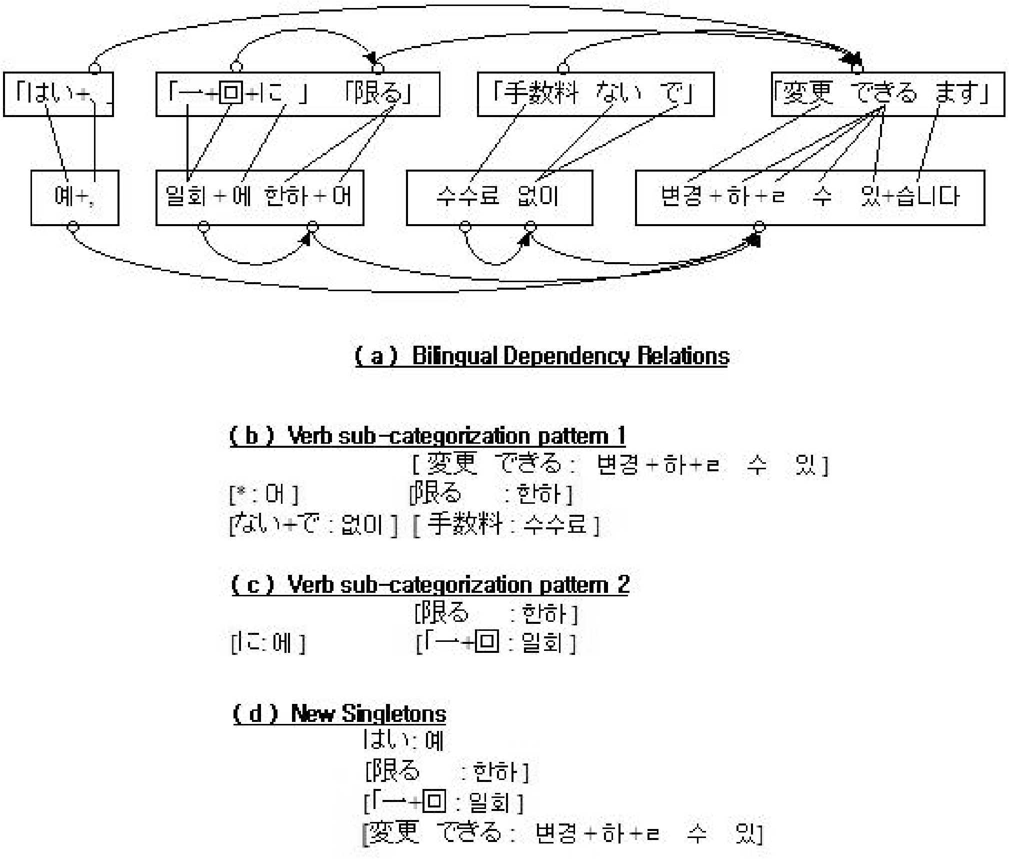


Figure 3: An example of JK bilingual knowledge extraction

sentence, and then choose immediate dependents of each predicate as an argument. If one of the immediate dependents is a predicate of a subordinate clause, then we also extract a verb sub-categorization pattern of a subordinate clause recursively. Figure 3 (b) shows a sentence predicate and its arguments, (c) a subordinate predicate and its arguments. [2]

However, some errors might result from dependency parsing, POS tagging, and alignment. In addition, the occurrence of a bilingual pair might happen accidentally. To obtain reliable ones, we count co-occurrences and use the χ^2 test. At first the χ^2 test is performed on a pair of translation correspondences which can be a pair of singletons or multi-word expressions. If it is over a given critical value, then we accept it as a reliable one. Otherwise, we reject it.

With reliable translation correspondences, we can construct bilingual translation knowledge bases such as surface verb sub-categorization patterns. A verb sub-categorization pattern is collocation of a verb and all of its argument/adjunct nouns. We represent a verb sub-categorization pattern by a feature structure *vsubcat* , which consists of a bilingual verb $Pred_j$: $Pred_k$ and all the pairs of co-occurring bilingual case markers $Func_j$: $Func_k$ and bilingual case marked nouns Arg_j : Arg_k:

[2] If an immediate dependent is an interjection, it is not regarded as an argument.

$$\mathbf{vsubcat} = \left[\begin{array}{ll} & Pred_j : Pred_k \\ Func_{j1} : Func_{k1} & Arg_{j1} : Arg_{k1} \\ . & \quad\quad\quad\ . \\ . & \quad\quad\quad\ . \\ . & \quad\quad\quad\ . \\ Func_{jn} : Func_{kl} & Arg_{jl} : Arg_{jn} \end{array}\right]$$

The verb sub-categorization pattern is decomposed into sub-frames $vcn_1 \cdots vcn_n$ which are composed of a verb and a pair of co-occurring case marker and case marked noun. Each sub-frame vcn_i is again decomposed into two subparts: one is a verb-case marker part, vc_i consisting of a verb and a case marker, the other is a verb-nominal part, vn_i consisting of a verb and a case marked noun. These decomposed sub-frames and subparts consist of a set of sub-frames, *VCN* and sets of subparts *VC* and *VN*, respectively.

$$\mathbf{vcn_i} = \ [Pred_j : Pred_k \quad Func_j : Func_k \quad Arg_j : Arg_k]$$

$$\mathbf{vc_i} = \ [Pred_j : Pred_k \quad Func_j : Func_k]$$

$$\mathbf{vn_i} = \ [Pred_j : Pred_k \quad Arg_j : Arg_k]$$

In order to get reliable bilingual knowledge base of surface verb sub-categorization patterns, we take the following stepwise filtering method:

- step 1. filter out the reliable translation correspondences from all of the alignment pairs by χ^2 test at a probability level of α_1
- step 2. decompose the extracted verb sub-categorization pattern into sub-frames, and a sub-frame into two sub-parts such as a verb-case marker part and a verb-nominal part
- step 3. filter out reliable sub-parts *vc* and *vn* by χ^2 test at a probability level of α_2
- step 4. filter out a reliable sub-frame *vcn* by unifying the reliable sub-parts, *vc* and *vn*
- step 5. filter out reliable verb sub-categorization patterns *vsubcat* through the unification of reliable sub-frames $vcn_1 \cdots vcn_n$

That is, we filter reliable sub-parts by χ^2 test on *vc* and *vn* where bilingual verb pairs and nominal pairs should be ones passed the χ^2 test of translation correspondences. Of a given sub-categorization sub-frames and patterns, we try to merge its sub-parts *vc* and *vn* into a sub-frame *vcn* and reliable sub-frames $vcn_1 \cdots vcn_n$ into a verb sub-categorization pattern *vsubcat* by applying the unification operation:

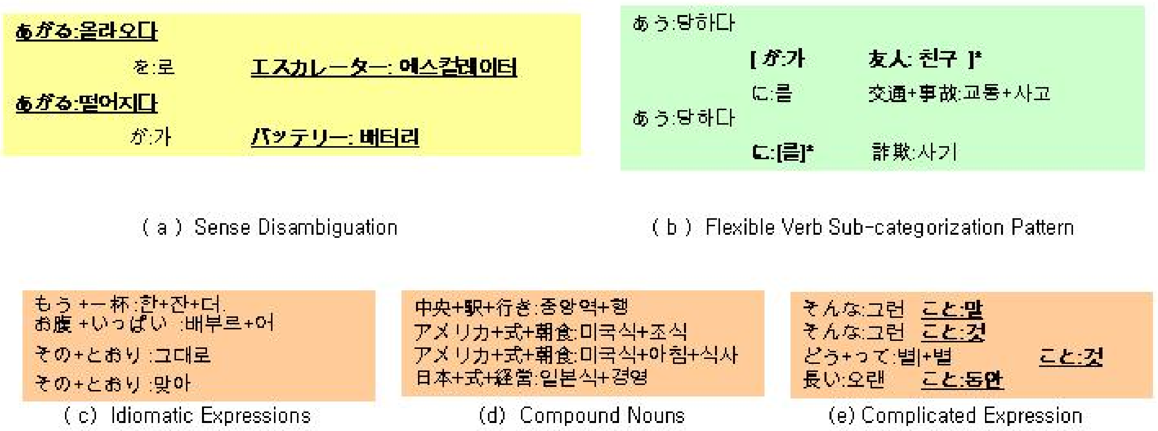
$$vc \wedge vn \rightarrow vcn \text{ s.t. } vc \in VC, vn \in VN \text{ and } vcn \in VCN$$

$$\wedge_{i=0}^{n} vcn_i \rightarrow vsubcat \text{ s.t. } vcn_i \in VCN \text{ and } vsubcat \in VSUBCAT$$

We assume that two sub-parts are unifiable into a sub-frame only if both of the sub-parts are reliable ones filtered by the χ^2 test and they share the same verb pair $< Pred_j : Pred_k >$ and judge it as plausible one. Of all the sub-frames of a given verb sub-categorization pattern, we apply the similar unification operation. If all the sub-frames of a given sub-categorization pattern can be merged by the unification operation, then we decide the verb sub-categorization patterns are reliable.

In addition, if each aligned sentence is composed of one chunk, which is composed of no less than three morphemes, we extract it as reliable translation knowledge.

Figure 4 shows some examples extracted by applying the proposed method.



(a) Sense Disambiguation

(b) Flexible Verb Sub-categorization Pattern

(c) Idiomatic Expressions

(d) Compound Nouns

(e) Complicated Expression

Figure 4: Examples of Sub-categorization Pattern and Multi-word Expression

	Training		Test	
	Japanese	Korean	Japanese	Korean
# of sentences	162,320		10,150	
# of total morphemes	1,153,954	1,179,753	74,366	76,540
# of bunsetsu/eojeol	448,438	587,503	28,882	38,386
vocabulary size	15,682	15,726	5,144	4,594

Table 1: Basic Travel Expression Corpus

3 Experiments

We used the Basic Travel Expression Corpus (BTEC), a collection of conversational travel phrases for Japanese and Korean (see Table 1). We used 162,320 sentences in parallel corpus for training and 10,150 sentences for test. The Japanese sentences were automatically dependency parsed by CaboCha[3], and the Korean sentences were automatically POS tagged with KUTagger(Rim,2003)

Through some experiments, we investigated the effect of dictionary lookup for alignment and evaluated the performance of the alignment by measuring the coverage and the average ambiguity of the extracted knowledge on the train and the test data set.

3.1 The Effect of Dictionary Lookup as a Preprocessing of Chunk/Word Alignment

Before chunk alignment, we referred to a bilingual dictionary and obtained 42.1% and 38.9% of aligned morphemes of Japanese and Korean, respectively. Table 2 shows the results. Especially, the ratios of the matched particles and verbs are very high. It can be interpreted as that 23.3% of Japanese bunsetsus were matched with Korean eojeols by particles, 15.5% by verbs and 38.8% by particles and verbs. In addition, if we include the other matched morphemes, ratio of Japanese bunsetsus aligned by only the dictionary lookup is over 45%. The high initial alignment lightened the burden of next chunk alignment and made the chunk alignment more accurate.

[3] This software is available at http://chasen.org/ taku/software/cabocha/

	translation pairs	uniq words	matched morph	matched particles	matched verbs
Japanese	101,590	40,481	485,920(42.1%)	104,700(23.3%)	69,445(15.5%)
Korean	101,590	34,479	506,836(38.9%)	104,700(18.7%)	69,445(15.5%)

Table 2: Coverage of Translation Dictionary Lookup on Training Set

	vc	vn	vcn	vsubcat
all	89.35(100)	70.87(100)	62.50(91.81)	56.25(83.8)
ave-ambiguities	28.59(25.77)	3.43(2.81)	2.81(2.33)	2.73(2.25)
# of bi-knowledge	102,114	121,700	111,784	65,053
# of uniq verbs				(23,431 13,382 10,789)
$\alpha 1 = 0.05, \alpha 2 = 0.1$	84.09(93.25)	64.86(87.22)	54.21(73.96)	47.78(66.27)
ave-ambiguities	19.45(17.77)	2.99(2.56)	2.45(2.14)	2.37(2.07)
# of bi-knowledge	48,163	83,280	61,422	41,807
# of uniq verbs				(13,254 8,350 6,835)
$\alpha 1 = 0.05, \alpha 2 = 0.05$	83.78(92.88)	64.64(89.91)	53.91(73.34)	47.48(65.70)
ave-ambiguities	19.07(17.44)	2.97(2.54)	2.44(2.14)	2.37(2.07)
# of bi-knowledge	46,815	82,595	60,188	41,150
# of uniq verbs				(13,072 8,450 6,741)
$\alpha 1 = 0.05, \alpha 2 = 0.01$	83.71(92.79)	64.60(86.80)	53.80(73.10)	47.37(65.49)
ave-ambiguities	18.00(16.48)	2.94(2.52)	2.43(2.13)	2.36(2.05)
# of bi-knowledge	46,174	82,157	59,712	40,833
# of uniq verbs				(12,916 8,417 6,713)

Table 3: Coverage of the Extracted and Filtered Bilingual Knowledge on Test and Training Set: (a) vsubcat(surface verb sub-categorization patterns),(b) vcn(tuples of bilingual verbs, case markers and nominals) (c) vn(pairs of bilingual verbs and nominals) (d) vc(pairs of bilingual verbs and case markers).

3.2 Evaluation of Alignment and Bilingual Knowledge Extraction

From the alignment results, we extracted verb sub-categorization patterns of 23,431 verb pairs of 13,382 Japanese and 10,789 Korean verbs. Through the stepwise filtering, we obtained reliable verb sub-categorization patterns and their sub-frames and sub-parts.

For evaluating the quality of the translation knowledge base and the alignment, we measured the coverage and average ambiguity of the bilingual knowledge extracted from the alignment results. In case of coverage, we checked whether there are corresponding verb sub-categorization patterns *vsubcat*, verb-case marker-nominal pairs *vcn*, verb-nominal pairs *vn* and verb-case marker pairs *vc* of a given sentence. Average ambiguity is measured as follows:

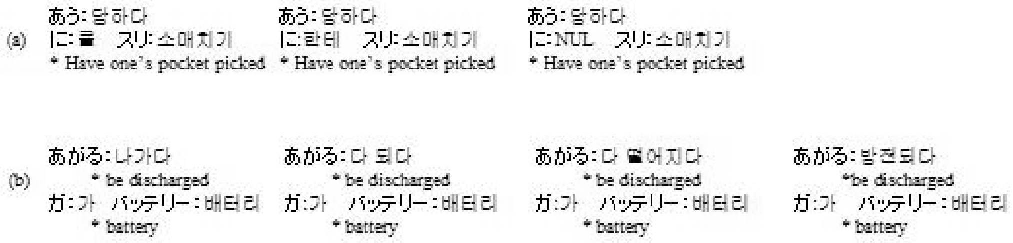
$$\text{Average ambiguity} = \frac{\sum_i \mid C_i \mid}{N} \tag{6}$$

where C_i is a set of translation pairs of a given translation target i, $|C_i|$ is the cardinality of the set and N is total number of translation targets in a sub-categorization, a sub-frame, and a sub-part level.

Table 3 shows the coverage, the average ambiguities, the number of filtered bilingual knowledge and the number of unique verb pairs, source verbs and target verbs according to the values of α_1 and α_2. In Table 3, values in parentheses denote the coverage and the average ambiguities on the training set. The numbers of unique pairs of Japanese and Korean verbs, Japanese verbs, and Korean Verbs are represented with a tuple of (verb pair, Japanese verb, Korean verb).

Even though we evaluated just only the coverage, we can approximate how many sentences are correctly aligned in a chunk and a word level by the coverage of vsubcat on the training and test set, how many chunks in the given sets are correctly aligned by the coverages of vcn, vn, and vc.

On the other hand, the coverage of the acquired sub-categorization knowledge is decreased as we increase the confidence level. If we check it with the average ambiguities, it is clear that the ambiguities in translation pairs are decreased and the bilingual knowledge is getting more reliable. Besides, if we consider the actual senses of translation pairs, the ambiguities in *vn*, *vcn* and *vsubcat* might be much lower because we regarded translation pairs with different surface forms as different ones. Figure 5 shows some examples of which surface forms are different but the meanings are the same.



(a)
あう:당하다
に:를 スリ:소매치기
* Have one's pocket picked

あう:당하다
に:한테 スリ:소매치기
* Have one's pocket picked

あう:당하다
に:NUL スリ:소매치기
* Have one's pocket picked

(b)
あがる:나가다
* be discharged
ガ:가 バッテリー:배터리
* battery

あがる:다 되다
* be discharged
ガ:가 バッテリー:배터리
* battery

あがる:다 떨어지다
* be discharged
ガ:가 バッテリー:배터리
* battery

あがる:방전되다
*be discharged
ガ:가 バッテリー:배터리
* battery

Figure 5: Examples of which surface forms are different but the meanings are the same

To be sure how accurate the alignment and the acquired bilingual knowledge are, we should evaluate the accuracy. However, it needs too much labor-intensive works. In order to solve this problem, we will apply the acquired knowledge to such application system as a machine translation and evaluate its performance. We will leave this issue for the near future work.

Nevertheless, from the experimental results, we found that by using the proposed method, we can obtain various effects as follows:

- Polysemy problem can be resolved by extracting the knowledge from the bilingual dependency parses which chunk and word aligned corpus (Figure 4 (a), (e))
- By compiling sub-categorization patterns of the same verb pair, we can acquire flexible usages of verb sub-categorization patterns(Figure 4 (b).
- Bilingual multi-word expression including idiomatic expressions and compound nouns can be easily extracted(Figure 4 (c), (d), (e)).

4 Related Works and Discussions

Over the last decade, research on machine translation has focused on automatic acquisition of knowledge from bilingual corpus, such as bilingual dictionaries, transfer rules and statistical models. Before automatically acquiring the knowledge, a number of issues remain to be addressed. Achieving accurate alignment is one of the fundamental problems. Alignment must produce the transfer knowledge that provide sufficient context to enable the translation system utilizing the knowledge to choose an appropriate translation for a given context.

To tackle the problem, structural or syntactic aspects of language are modeled. Some approaches, which aims at applying statistical models to structural data, have begun to emerge(Yamada et al,2001)(Ding et al,2003)(Gildea,2003). They tried to align hierarchical structures like sub-trees with parsing technologies. (Yamada et al,2001) statistically modeled the translation process from a parse tree of source language into a target language sentence. (Gildea,2003)(Ding et al,2003) tried alignment of tree-to-tree while allowing the tree structure to constrain the alignment at the high level, but relaxing the isomorphism constraints as necessary within smaller subtrees. (Imamura,2002) proposed the hierarchical phrase alignment to extract machine translation knowledge. For the alignment, they introduced the syntactic and semantic constraints with the translation dictionaries. However, as mentioned previously, parse-to-parse matching, which regards parsing and alignment as separate and successive procedures, suffers from grammatical inconsistency between languages. Moreover, it costs a lot to develop parsers of both the source and target language.

On the contrary to the previous research, we used a monolingual dependency parser to automatically obtain dependency parses of target language using chunk and word alignment. By sharing the dependency relations of a given source sentence, we could obtain a dependency parse of a target sentence that is structurally consistent with the source sentence. It enabled us to save the development cost of a parser for target language.

On the other hand, research for acquiring the sub-categorization knowledge have been focused on the monolingual sub-categorization(Utsuro et al,1993)(Kawahara et al,2001). There are few research for acquisition of bilingual sub-categorization (Fung et al,2004). (Fung et al,2004) tried to construct a bilingual semantic network, BiFrameNet to enhance statistical and transfer-based machine translation systems. They induced the mapping between the English lexical entries in FrameNet to Chinese word senses in HowNet. Their approach is a quite different from ours, in the sense of utilizing the existing resources. It takes such an advantage of generalized bilingual frame semantics while our bilingual knowledge is not generalized yet. However, they have still problems of appropriate mapping from lexical entries to word senses and obtaining correct example sentences. On the contrary, our strong point lies in automatically resolving the bilingual word senses via chunk and word alignment.

5 Conclusion

In this paper, we have shown a method of acquiring bilingual knowledge using chunk and word alignment. We applied such NLP technologies as dependency parsing and POS tagging to incorporate structural and syntactic aspects of languages. To maintain structural consistency between languages, we used a monolingual dependency parser to automatically obtain dependency parses of target language using chunk and word alignment. Moreover, we have shown that the method using a bilingual dictionary and adopting a divide-and-conquer strategy effectively reduced the computational complexity of structural alignment and also obtained high accurate alignment, which enabled the acquisition of more reliable translation knowledge.

In the near future, we will apply the acquired knowledge to machine translation and show the usefulness of them. Furthermore, we will try to cluster the verb sub-categorization patterns and construct generalized knowledge base of them by incorporating with existing thesauri.

References

Eiji Aramaki, Sadao Kurohashi, Satoshi Sato & Hideo Watanabe:2001, ‘ Finding translation correspondences from parallel parsed corpus for example-based translation’, in MT Summit VIII, Santiago de Compostela, Spain, pp.27-32.

Hiyan Alshawi, Srinivas Bangalore and Shona Douglas: 2000, ‘Learning Dependency Translation Models as Collections of Finite State Head Transducers,’ in *Computational Linguistics* 26(1): 45-60.

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

Yuan Ding, Daniel Gildea, and Martha Palmer: 2003, ‘ An Algorithm for Word-level Alignment of Parallel Dependency Trees’, in *The 9th Machine Translation Summit of the International Association for Machine Translation* , New Orleans

Pascale Fung and Benfeng Chen:2004, ‘ BiFrameNet: Bilingual Frame Semantics Resource Construction by Cross-lingual Inductio’, in *Proc. of The 20th International Conference on Computational Linguistics,(COLING 2004)*,Geneva, Switzerland

Daniel Gildea:2003, ‘ Loosely Tree-Based Alignment for Machine Translation’, in *Proceedings of the 41th Annual Conference of the Association for Computational Linguistics (ACL-03)* , Sapporo, Japan.

Kenji Imamura : 2002, ‘ Application of Translation Knowledge Acquired by Hierarchical Phrase Alignment for Pattern-based MT’, in *The 9th Conference on Theoretical and Methodological Issues in Machine Translation (TMI-2002)*, pp. 74-84.

Daisuke Kawahara and Sadao Kurohashi: 2001, ‘Japanese Case Frame Construction by Coupling the Verb and its Closest Case Component,’ In *Proc. of First International Conference on Human Language Technology Research (HLT 2001)*, pp.204-210, San Diego, California

Arul Menezes & Stephen D. Richardson: 2001, ‘ A best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora,’ in *Proceedings of MT Summit VIII, Santiago De Compostela*, Spain

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 *Proc. of LREC 2002*, pp. 147-152, Spain, May

Takehito Utsuro, Yuji Matsumoto and Makoto Nagao:1993, ‘Verbal Case Frame Acquisition from Bilingual Corpora,’ in *Proc. of IJCAI*, pp. 1150-1157

Hideo Watanabe, Sadao Kurohashi and Eiji Aramaki: 2000, ‘Finding Structural Correspondences from Bilingual Parsed Corpus for Corpus-based Translation,’ in *COLING 2000*, pp. 906-912

Kenji Yamada and Kevin Knight: 2001, ‘ A Syntax-based Statistical Translation Model,’ in *ACL 2001*, pp. 523-530

Hae-Chang Rim: 2003, ‘ Korean Morphological Analyzer and Part-of-Speech Tagger,’ *Technical Report,* NLP Lab. Dept. of Computer Science and Engineering, Korea University