Hierarchical Phrase Alignment Harmonized with Parsing

Kenji Imamura
ATR Spoken Language Translation Research Laboratories
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0288, Japan
kenji.imamura@atr.co.jp

Abstract

In this paper, we propose a hierarchical phrase alignment method that aims to acquire translation knowledge. Previous methods utilize the correspondence of sub-trees between bilingual parsing trees after determining the parsing result. The method described in this paper combines partial tree candidates, and selects the best sequence of partial trees. Then, a structural similarity measure (called a 'phrase score') is used for evaluation. A forward DP backward A^* search algorithm is applied in order to combine partial trees. Using this method, about twice as many as equivalent phrases were extracted experimentally, and almost no deterioration was observed.

In addition, we found that a word alignment function with a high recall rate is suitable for this method.

1 Introduction

When building a machine translation system, other than statistical translation systems that apply bilingual corpora directly as translation knowledge, we have to construct knowledge such as translation rules manually. Therefore, automatic translation knowledge construction is an effective way to reduce costs when we apply the system to other domains.

In this paper, we propose a hierarchical phrase alignment method that aims to acquire translation knowledge automatically from bilingual sentences. Here, phrase alignment (PA) refers to the extraction of equivalent partial word sequences between bilingual sentences. We use the term phrase alignment since these word sequences include not only words but also noun phrases, verb phrases, relative clauses, and so on. English and Japanese languages are used in this study.

For example, in the case of the following sentence pair,

E: *I have just arrived in New York.*
J: *NewYork ni tsui ta bakari desu.*

the phrase alignment method should extract the following word sequence pairs.

- *in New York* ↔ *NewYork ni*
- *arrived in New York* ↔ *NewYork ni tsui*
- *have just arrived in New York* ↔ *NewYork ni tsui ta bakari desu*

We call these 'equivalent phrases' in this paper.

Equivalent phrases indicate corresponding expressions between two languages, so that they can be applied to example-based translation systems directly. In addition, because the phrases maintain hierarchical information, translation knowledge can be compressed by making hierarchical patterns.

Some phrase alignment methods have previously been proposed, such as those of Kaji et al. (1992), Matsumoto et al. (1993), Kitamura and Matsumoto (1995), Yamamoto and Matsumoto (2000), and Meyers et al. (1996). The characteristics common to these previous methods are:

1. The methods employ parsers (phrase structure analyzers or dependency analyzers) and word alignment (WA) results.
2. When they search for phrase correspondences, they only handle the final structures that the parser has output.
3. They handle only content word correspondences.

However, in the previous methods, the phrase alignment results directly depend on the parsing accuracy.

To cope with this problem, we propose a new phrase alignment method that is harmonized with parsing. In addition, we increase PA accuracy by

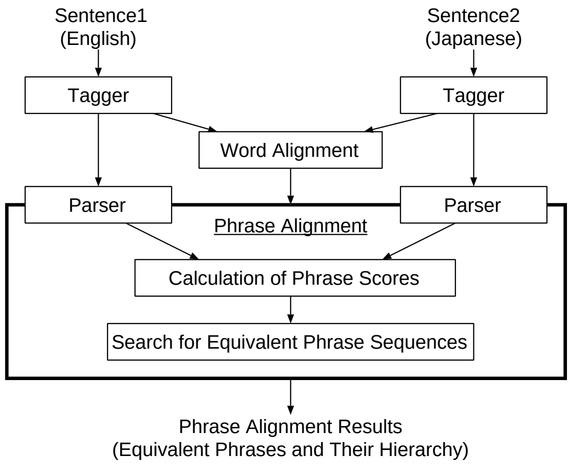

Figure 1: Flow of Hierarchical Phrase Alignment

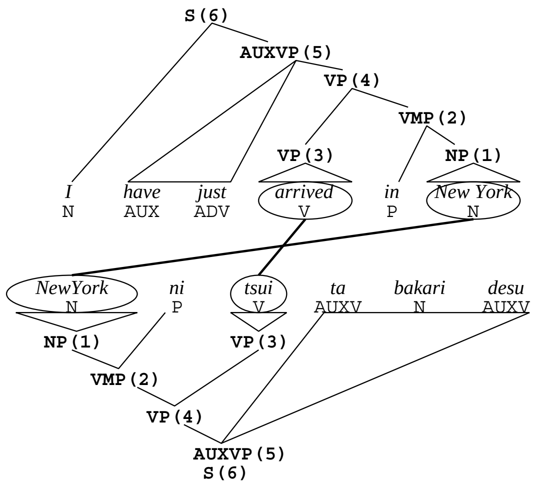

Figure 2: An Example of Simple Translation

using as WA results not only content words but also functional words.

In the next section, we explain the basic method of hierarchical phrase alignment. Section 3 tells how to harmonize the basic method with parsing, Section 4 describes suitable functions of word alignment for the phrase alignment, and Section 5 evaluates the performance, including comparisons with alternative methods.

2 The Basic Method of Hierarchical Phrase Alignment

From a local view, most phrases in manual translation are translated into phrases of the same type, even if the language families are different. For example, the English verb phrase "*arrive in New York*" is generally translated into the Japanese verb phrase "*NewYork ni tsuku*".

Considering this feature, we assume that if the word sequences of bilingual sentences have the same semantic information, and if the phrase types are equal, the sequences can be regarded as equivalent phrases. We interpret this assumption as follows for computing:

Condition 1: "The same semantic information" → "Words in the pair corresponded with no deficiency and no excess"

Condition 2: "The same phrase types" → "The phrases are of the same syntactic category"

In order to extract phrases that satisfy the above two conditions, the flow is shown as follows (Figure 1).

1. Tag and parse an English sentence and a Japanese sentence.
2. Extract corresponding words (called 'word links') by word alignment. We assume that W word links are extracted. Since many word alignment methods have been proposed elsewhere [1], we do not discuss how they work in this paper.
3. Select i word links from among all of the links ($1 < i \leq W$), catch all of the syntactic nodes (non-terminal symbols) that include the links, and exclude all other word links in the leaves from the parsed English tree and Japanese tree.
4. Compare the syntactic categories of all English and Japanese nodes captured in process 3. When identical node categories are found, regard the leaves of the nodes as equivalent phrases. If candidates of a sentence or auxiliary verb phrase category are acquired, the candidate which covers the maximum area is selected. In other ambiguous cases, the candidate which covers the minimum area is selected.
5. Try processes 3. and 4. for all word link combinations.

Example (1): For example, we consider the English sentence "*I have just arrived in New York.*" and its translation "*NewYork ni tsui ta bakari desu*", which have two word links, i.e., between '*New York*' and '*NewYork*' and between '*arrive*' and '*tsui*'. Given the parsing trees and word links in Figure 2, the equivalent phrases are extracted as follows.

[1] e.g., Melamed (2000) and Sumita (2000)

Table 2: The Type of Syntactic Categories

Phrase Type	Mark
Noun Phrase	NP
Verb Phrase	VP
VP with Auxiliary Verbs	AUXVP
Verb Modifier Phrase	VMP
Noun Modifier Phrase	NMP
Indepenent Phrase	INDP
Sentence	S
Other (Language Dependent Phrase)	

1. The syntactic node pair that consists only of the '*New York*' – '*NewYork*' word link (i.e., exclude the '*arrived*' – '*tsui*' link) and that contains phrases of the same syntactic category is retrieved. This finds the phrases `NP(1)` and `VMP(2)`.
2. Next, the syntactic node pair that consists only of the '*arrived*' – '*tsui*' word link (i.e., exclude the '*New York*' – '*NewYork*' link) and that contains phrases of the same syntactic category is retrieved. This finds the phrase `VP(3)`.
3. Finally, the node pairs that include both the '*New York*' – '*NewYork*' link and the '*arrived*' – '*tsui*' link and are of the same category are retrieved. This finds the phrases `VP(4)`, `AUXVP(5)`, and `S(6)`.

Therefore, the six equivalent phrases shown in Table 1 are extracted.

This is an example of two word links. In the case of three word links, the method retrieves phrases that include combinations of word links, such as those including link 1, including links 1 and 2, and including all links. Equivalent phrases are extracted hierarchically.

Since the syntactic categories are different between English and Japanese, we classified the categories into seven types, as shown in Table 2, which are common to both languages. Using this classification, we are able to compare different language categories.

Example (2): Even though a word link is available, Part-Of-Speech of words are often different in different languages. If corresponding phrases are sought for using such word links with no syntactic constraint, inappropriate translations will be extracted. However, the method in this paper acquires only syntactically similar phrases, so that a few unnatural short phrases are extracted as equivalents.

For example, consider extracting equivalent phrases from the English sentence "*Business class is fully booked.*" and the Japanese sentence "*business-class ha yoyaku de ippai desu*" (Figure 3). Even if the word alignment links '*fully*/ADV' – '*ippai*/N', and '*booked*/V' – '*yoyaku*/N' are given, there are no nodes that have a link between them and are of the same syntactic category. However, there are `VP(2)` nodes that both include links and are of the same category, so the English phrase "*be fully booked*" and the Japanese phrase "*yoyaku de ippai desu*" are regarded as equivalents.

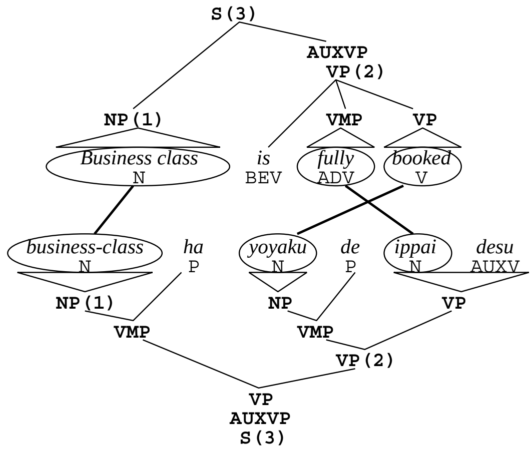

Figure 3: An Example of Different POS Word Links

Example (3): An example of a non-literal translation is shown in Figure 4. In this example, because the English phrase "*fly*" is translated into the Japanese phrase "*hikoki de yuki* (go by plane)", they have no word links. However, the final result of phrase alignment contains the English phrase "*fly to New York tomorrow*" and the Japanese phrase "*NewYork ni asu hikoki de yuki*" as equivalents, so that "*fly*" and "*hikoki de yuki*" are indirectly regarded as equivalents. Thus, this method is able to extract some non-literal equivalent phrases (i.e., non-word-by-word translation phrases) that lack word links.

Phrase alignment with a lack of word links is described in detail in Section 4.2.

3 Phrase Alignment Harmonized with Parsing

The basic method described in Section 2 assumes that single parsing trees are given. However, phrase alignment results are directly affected by parsing results when they are processed after the determination of a single parsing tree. For example, a sentence pair in which the parser cannot analyze their structure cannot be used to process

Table 1: An Example of Phrase Alignment Results

Syn. Category	English Phrase	Japanese Phrase
NP	*New York*	*NewYork*
VMP	*in New York*	*NewYork ni*
VP	*arrive*	*tsuku*
VP	*arrive in New York*	*NewYork ni tsuku*
AUXVP	*have just arrived in New York*	*NewYork ni tsui ta bakari desu*
S	*I have just arrived in New York*	*NewYork ni tsui ta bakari desu*

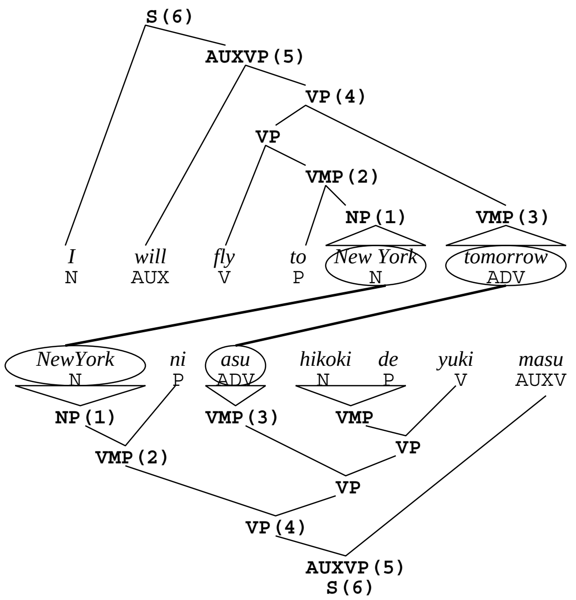

Figure 4: An Example of a Non-literal Translation

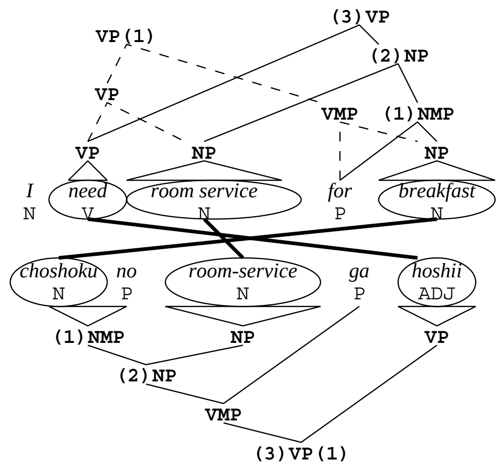

Figure 5: An Example of Disambiguation for a PP Attachment Modifyee

phrase alignment. Moreover, incorrect parsing trees derive incorrect or insufficient phrase alignment results.

Parsing errors can be roughly classified into two types. One is when the parsing result contains ambiguities and the parser selects the wrong candidate. The other is when the parsing process fails because of incomplete grammar (i.e., a lack of rewrite rules). Our method solves these problems by harmonizing the parsing by using the following two features and techniques.

3.1 Disambiguation Using Structural Similarity between Languages

Some parsing ambiguities can be eliminated when the two languages are made to correspond. This disambiguation utilizes structural similarity (Kaji et al., 1992).

For example, the PP attachment modifyee in English is itself disambiguous when the equivalent Japanese phrase has only one structure. In Figure 5, the prepositional phrase "*for breakfast*" may modify either '*need*' and consist of `VP(1)` shown by the dotted line tree, or it may modify '*room service*' and consists of `(2)NP` shown by the solid line tree. On the other hand, considering the Japanese sentence structure, "*choshoku no*" decisively modifies '*room-service*' and consist of `(2)NP`. Therefore, "*room service for breakfast*" must be a noun phrase in the same way as Japanese.

This phenomenon indicates that the situation of ambiguities depends on the language, and some ambiguities can be eliminated by intersecting their situations.

Thus, an evaluation measure that outputs a high score when the structures of two languages become more similar is able to achieve some disambiguation.

We set the evaluation measure as follows:

- Form a correspondence for all English and Japanese nodes with the two conditions described in Section 2.
- And select the structure that has the maximum number of corresponding nodes.

We call this measure **'phrase score'** in this paper. In the case of the solid line structure in Figure 5, `(1)NMP`, `(2)NP`, and `(3)VP` are evaluated as corresponding nodes. In the case of the dotted line structure, only VP(1) is evaluated as a corresponding node in the same area. Therefore, the phrase score of the solid line structure is 2 greater than that of the dotted one, and the solid structure is selected.

Note that we assumed that there are no ambiguities in the word alignment result. If there are ambiguous word links (e.g., the same word appears twice in a sentence), the above evaluation measure can disambiguate them to some degree by searching for word link combinations that maximize the phrase score.

3.2 Combination of Partial Trees

The phrase alignment method in this paper uses a chart parser. Most parsers, including this one, output nothing when they fail to construct a parsing tree due to incomplete grammar (i.e., a lack of rewrite rules), nevertheless they still keep partial trees in their agenda. Namely, correct partial tree candidates remain in the agenda. If we combine these partial trees appropriately, we can recover the failure caused by incomplete grammar. This approach is especially effective for spoken language that has a lot of ungrammatical sentences (Takezawa and Morimoto, 1997). The evaluation measure that we described in Section 3.1 is useful for the examining this.

Naturally, when the parsing is successful (i.e., a tree is derived from a whole sentence), the result should be preferred. Thus, we revise the measure so that the result of the minimum number of partial trees will be preferred. The evaluation measure is finally represented as follows.

1. Compare the nodes of a sentence pair, and extract the equivalent phrase candidates that maximize the phrase score.
2. Calculate the sum of the phrase scores in the partial tree sequence, and select the sequences that have the maximum score.
3. If plural partial tree sequences are available, select the sequences that have the minimum number of partial trees.

However, exponential time is necessary to examine all of the combinations of partial trees. In order to avoid this problem, we employ a 'forward DP backward A^* algorithm (Nagata, 1994)', which is a two-pass search algorithm used for taggers.

The algorithm is briefly described as follows:

First, all partial trees (the phrase scores of partial trees have already been calculated) are constructed into a lattice structure.

Second, on the forward search, the maximum score, called the estimation score, which is a total of the phrase scores through the sequence, is calculated using dynamic programming from the start of the sentence to edge $i (0 \leq i \leq \text{word number} N)$. Note that the paths are not recorded at this time. The estimation score means that there is at least one path in the score from the start to edge i.

On the backward search, the best combination of partial trees is searched for by using the $A*$ search algorithm. The estimation score is used as the heuristic function value of the $A*$ algorithm at that time. Because the estimation score is the most accurate value of the heuristic function, the best path is searched for but no redundant paths are expanded while searching.

Thus, the best partial tree sequence (i.e., the best equivalent phrase sequence) can be extracted by using this search algorithm without pruning, and the searching time is almost in proportion to the word number.

4 Word Alignment for Phrase Alignment

4.1 Correspondence between Functional Words and Content Words

Functional words represent aspects, moods and so on, and they expand the variety of expressions. If we were to extract equivalent phrases while ignoring the functional words, phrases which are correct in the view of meaning but incorrect in the view of pragmatics would be extracted as the translation.

Figure 6 shows an example. If there is no word link between '*after*' and '*iko*', NP(1) is extracted as an equivalent because there is only a word link between '*three*' and '*sanji*' in the phrase. However, if there is a word link between '*after*' and '*iko*', NP(1) is ignored because there is a deficient/excess link.

Word links between functional words or between a functional word and a content word make the constraint of Condition 1 tighter. Therefore, incorrect equivalent phrases might be ignored, and consequently, the accuracy of phrase alignment will be increased.

4.2 Relationship between Word Alignment Accuracy and Phrase Alignment

No word alignment (WA) methods that have 100% precision and recall rates have been proposed. Therefore, we should consider that word links include word alignment errors. With this in mind, which rate is more important for phrase alignment, precision or recall?

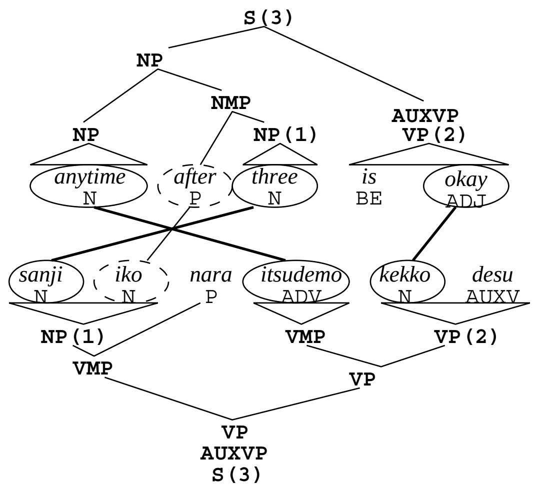

Figure 6: An Example of a Functional Word Link

Under the condition of a 100% WA recall rate, a low WA precision rate means that the word links include redundancy. As we described in Section 4.1, if the word link number is increased, Condition 1 becomes tighter. Therefore, the number of equivalent phrases will be decreased, but a few incorrect equivalent phrases will be extracted. [2]

On the other hand, in the case of a low WA recall rate (i.e., word links are insufficient), Condition 1 becomes looser, and ambiguities, such as PP attachment modifyees increase. Therefore, incorrect equivalent phrases will be extracted. In addition, the number of equivalent phrases is decreased along with the decrease in the number of word link combinations.

Briefly put, a word alignment method that has a high recall rate is more suitable for phrase alignment. In other words, word links should include most of the necessary links even though they include redundancy.

5 Trial Experiments

5.1 Experiment Conditions

We conducted a number of experiments for phrase alignment. Three hundred bilingual sentences containing basic travel expressions were used.

The details of the experimental conditions are as follows:

- Tagging data was prepared by machine tagging with manual correction.
- Word links between content words were made manually. Links between functional words were made by referring to a translation dictionary.
- A basic bottom-up chart parser was used. The grammar was CFG that contained 286 English rules and 254 Japanese rules. The properties of the parser are shown in Table 3. The accuracy was 44% for the English test set, and 52% for the Japanese test set, which is lower than other parsers. For example, the accuracy of the Apple Pie Parser (Sekine and Grishman, 1995)[3] is over 70%.
- In each experiment, three hundred phrases were selected from the first candidate of the phrase alignment results, and bilingual evaluators evaluated them. One evaluator was a native English speaker, and another was a native Japanese speaker. The following three ranks were used for evaluation.

  A: Correct. It was a possible translation in the view of English to Japanese and Japanese to English.

  B: Not wrong, but depends on context. It was a possible translation in the view of English to Japanese or Japanese to English.

  C: Incorrect. It was a wrong translation in both English to Japanese and Japanese to English.

Table 3: The Properties of the Parser and the Grammar

Property	English	Japanese
Sentence#	300	300
Output Sentence# (rate)	200 (67%)	200 (67%)
Total Candidate# (per Output)	836 (4.18)	394 (1.97)
The 1st Candidate Accuracy per Sentence (per Output)	44% (67%)	52% (78%)

5.2 Effects of Harmonization

First, we tested a variety of phrase alignment methods for the following three cases. The statistics of the phrase alignment results are shown in Table 4.

Case 1: (Proposed Method) Extracting partial trees from the agenda while parsing, and searching for the best sequence with the maximum phrase score.

[2] Incorrect equivalent phrases are unfortunately extracted when there are redundant word links that satisfy Condition 2.

[3] http://cs.nyu.edu/cs/projects/proteus/sekine

Case 2: Selecting the first parsing candidate only when whole sentence has been parsed, and processing the phrase alignment.

Case 3: Using all of the parsing candidates only when the whole sentence has been parsed, processing the phrase alignment for all combinations of candidates, and selecting the best one with the maximum phrase score. Compared with Case 2, the effect of the phrase score became clearer. Moreover, compared with Case 1, the effect of the partial tree combination became clearer.

First, in Case 1, the equivalent phrase accuracy (we only consider Rank A) was about 86%. When we consider parsing accuracy, phrase alignment accuracy is higher when individual parsing trees are compared.

Comparing Cases 2 and 3, the number of equivalent phrases was increased in Case 3. However, the accuracy of equivalent phrases was almost the same. The reason for the same accuracy is that this method essentially makes few incorrect correspondences, so that incorrect parsing trees were ignored when extracting equivalent phrases.

Comparing Cases 3 and 1, almost all sentences were analyzed, some equivalent phrases were output in Case 1, and the number of equivalent phrases became about double. Moreover, the accuracy was almost the same. Therefore, the phrase scores work appropriately during a combination of partial trees, and the combination is especially effective for spoken language.

5.3 The Influence of Word Alignment

In order to examine the influence of word alignment, we tested using various word links. All of the experiments used the Case 1 phrase alignment method. The results are shown in Table 4.

Case 4: Word links are limited to content words. Compared with Case 1, the effect of the functional word links became clearer.

Case 5: A variable WA precision rate under a fixed WA recall rate. The word links in Case 1 were regarded as perfect, and the precision rate was changed from 50% to 100%. The purpose of this case is to measure the influence of redundant word links. The redundant word links were made by selecting word pairs randomly that are not included in the original links.

WA precision and recall rates are represented by the following equations:

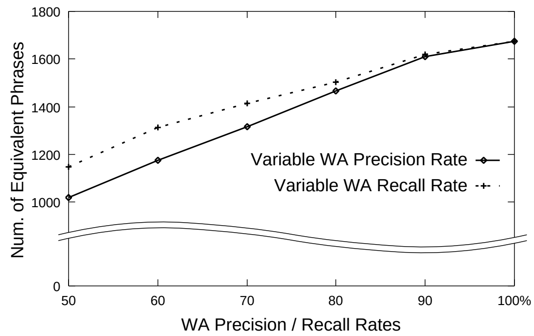

Figure 7: The Number of Extracted Equivalent Phrases According to Word Alignment Accuracies

$$\text{WA precision rate} = \frac{N_{org}}{N_{org} + N_{red}}$$

$$\text{WA recall rate} = \frac{N_{org} - N_{eli}}{N_{org}}$$

Where N_{org} is the number of original word links, N_{red} is the number of redundant links, and N_{eli} is the number of eliminated word links.

Case 6: A variable WA recall rate under a fixed WA precision rate. The recall rate was changed from 50% to 100%. The purpose of this case is to measure the influence of insufficient word links. The insufficient links were made by eliminating links randomly.

The Effect of Function Word Links: Comparing Cases 1 and 4, the number of extracted equivalent phrases was increased slightly in Case 4, and the accuracy of the equivalent phrases decreased slightly. By checking the correctness of different phrases between Cases 1 and 4, we confirmed that the phrases in Case 1 were more correct. Thus, there is no doubt that function word links are necessary for phrase alignment.

Influence of Word Alignment Accuracy: Figure 7 shows the number of extracted equivalent phrases in Cases 5 and 6 according to variable WA precision and recall rates. In both cases, the number of phrases decreased similarly according to the decrease in WA accuracy, but this was slightly affected by the WA precision rate.

On the other hand, Table 4 indicates that when the WA precision rate decreased, the accuracy of the equivalent phrases was almost equal, but when the WA recall rate decreased, the accuracy clearly decreased. Therefore, as we described in Section 4.2, the WA recall rate affects the phrase alignment accuracy with greater sensitivity. In other words, it is easier to increase the phrase alignment

Table 4: Number of Equivalent Phrases and Their Accuracy

	Case	Output Sentence#	Equivalent Phrase# (per Output)	Accuracy of Equivalent Phrase		
				Rank	Phrase#	Rate
Proposed Method	Case 1	296	1,676 (5.66)	A	248 + 269	86.2%
				B	30 + 5	5.8%
				C	22 + 26	8.0%
Alternative Phrase Alignment Methods	Case 2	176	726 (4.13)	A	249 + 270	86.5%
				B	30 + 8	6.3%
				C	21 + 21	7.0%
	Case3	177	822 (4.64)	A	264 + 267	88.5%
				B	18 + 3	3.5%
				C	18 + 30	8.0%
Variable Word Links	Case 4	295	1,703 (5.77)	A	240 + 258	83.0%
				B	31 + 4	5.8%
				C	29 + 36	10.8%
	Case 5 WA Precision: 50% WA Recall: 100%	276	1,018 (3.69)	A	245 + 266	85.2%
				B	17 + 0	2.8%
				C	38 + 31	11.5%
	Case 6 WA Precision: 100% WA Recall: 50%	272	1,147 (4.22)	A	209 + 230	73.2%
				B	21 + 4	4.2%
				C	70 + 66	22.7%

accuracy by providing as many word links as possible.

6 Conclusion

In this paper, we proposed a hierarchical phrase alignment method harmonized with parsing. The method uses a 'phrase score' to evaluate syntactic structural similarity, and carries out disambiguation and partial tree combination.

In particular, the proposed method can extract about double the test set in comparison with independent parsing. The accuracy of the extracted equivalent phrases was about 86%, and almost no deterioration was observed.

Since the proposed method has greater sensitivity to a lack of necessary word links, it extracts better equivalent phrases when it uses a word alignment method with a high recall rate.

We are planning to acquire translation knowledge from the extracted equivalent phrases, and apply it to an example-based translation system.

References

Hiroyuki Kaji, Yuuko Kida, and Yasutsugu Morimoto. 1992. Learning translation templates from bilingual text. In *Proceedings of COLING-92*, pages 672–678.

Mihoko Kitamura and Yuji Matsumoto. 1995. A machine translation system based on translation rules acquired from parallel corpora. In *Proceedings of Recent Advances in Natural Language Processing*, pages 27–36.

Yuji Matsumoto, Hiroyuki Ishimoto, and Takehito Utsuro. 1993. Structural matching of parallel texts. In *the 31st Annual Meeting of the ACL*, pages 23–30.

I. Dan Melamed. 2000. Models of translational equivalence among words. *Computational Linguistics*, 26(2):221–249, June.

Adam Meyers, Roman Yangarber, and Ralph Grishman. 1996. Alignment of shared forests for bilingual corpora. In *Proceedings of COLING-96*, pages 460–465.

Masaaki Nagata. 1994. A stochastic Japanese morphological analyzer using a forward-dp backward-A* N-best search algorithm. In *Proceedings of COLING-94*, pages 201–207.

Satoshi Sekine and Ralph Grishman. 1995. A corpus-based probabilistic grammar with only two non-terminals. In *the Fourth International Workshop on Parsing Technology, Prague, Czech.*

Eiichiro Sumita. 2000. Word alignment using a matrix. In *Proceedings of PRICAI2000*, page 821. Springer.

Toshiyuki Takezawa and Tsuyoshi Morimoto. 1997. Dialogue speech recognition using syntactic rules based on subtrees and preterminal bigrams. *Systems and Computers in Japan*, 28(5):22–32.

Kaoru Yamamoto and Yuji Matsumoto. 2000. Acquisition of phrase-level bilingual correspondence using dependency structure. In *Proceedings of COLING-2000*, pages 933–939.