# Automatic Acquisition of Paraphrases Using Bilingual Dependency Relations

Young-Sook Hwang and Young-Kil Kim

*ABSTRACT—This letter introduces a new method to automatically acquire paraphrases using bilingual corpora. It utilizes the bilingual dependency relations obtained by projecting a monolingual dependency parse onto the other language's sentence based on statistical alignment techniques. Since the proposed paraphrasing method can clearly disambiguate the sense of the original phrases using the bilingual context of dependency relations, it would be possible to obtain interchangeable paraphrases under a given context. Through experiments with parallel corpora of Korean and English language pairs, we demonstrate that our method effectively extracts paraphrases with high precision, achieving success rates of 94.3% and 84.6%, respectively, for Korean and English.*

*Keywords—Paraphrase, bilingual dependency parsing, alignment, sense disambiguation, dependency relation.*

## I. Introduction

Approaches based on bilingual corpora are promising for the automatic acquisition of translation knowledge. Phrase-based statistical machine translation (SMT) models have advanced the state of the art in machine translation by expanding the basic unit from words to phrases [1], [2]. However, phrase-based SMT techniques suffer from data sparseness problems, such as unreliable translation probabilities of low-frequency phrases and low coverage, in that many phrases encountered at run-time are not observed in the training data. An alternative to these problems is to use paraphrases.

In this study, we introduce a method of automatically acquiring paraphrases to smooth the translation parameters and to increase the coverage of translation knowledge. One previous approach identifies paraphrases using a phrase in another language as a pivot without contextual information [3]. Unlike the previous approach, our method takes the bilingual context into account in order to disambiguate the sense of paraphrases. First, we collect the bilingual dependency relations aligned with the same dependency relation in the other language as paraphrased dependency relations. Then, we extract the phrases sharing the same head (or modifier) phrase in the set of the paraphrased dependency relations aligned with a unique dependency relation in the other language. They are conceptually equivalent paraphrases.

This work is based on two converse assumptions of word sense. If multiple phrases map onto a single foreign language phrase, then the senses of the phrases are similar. On the other hand, if a single phrase maps onto different phrases in the foreign language, the senses of different phrases might be different. The two-step paraphrasing method allows us to increase the precision of the paraphrases by constraining the paraphrase candidates under the bilingual contexts of dependency relations.

To systematically acquire paraphrases, our method includes the following steps. First, we derive a bilingually parsed sentence by projecting the source language parse onto the word- and phrase-aligned target sentence. Second, we extract phrase correspondences and bilingual dependency relations from the bilingual dependency parses. Third, we acquire paraphrases by exploiting the extracted correspondences and dependency relations.

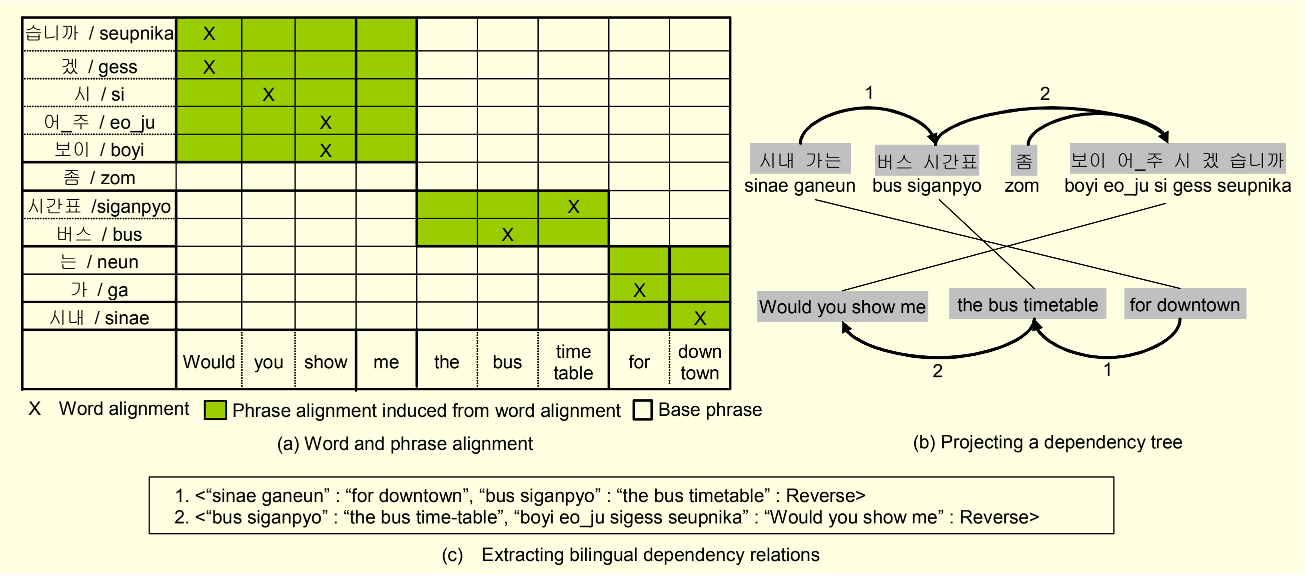
## II. Acquisition of Paraphrases

### 1. Extracting Bilingual Dependency Relations

To acquire bilingual dependency relations, we use word/phrase alignment techniques and the bilingual dependency parsing technique of projecting a dependency parse onto a

Manuscript received Aug. 31, 2007; revised Oct. 31, 2007.

Young-Sook Hwang (phone: + 82 42 860 6475, email: yshwang7@etri.re.kr) and Young-Kil Kim (email: kimyk@etri.re.kr) are with the Embedded SW Research Division, ETRI, Daejeon, Rep. of Korea.



| | Would | you | show | me | the | bus | time table | for | down town |
|---|---|---|---|---|---|---|---|---|---|
| 습니까 / seupnika | 🟩 X | 🟩 | 🟩 | 🟩 | | | | | |
| 겠 / gess | 🟩 X | 🟩 | 🟩 | 🟩 | | | | | |
| 시 / si | 🟩 | 🟩 X | 🟩 | 🟩 | | | | | |
| 어_주 / eo_ju | 🟩 | 🟩 | 🟩 X | 🟩 | | | | | |
| 보이 / boyi | 🟩 | 🟩 | 🟩 X | 🟩 | | | | | |
| 좀 / zom | | | | | | | | | |
| 시간표 /siganpyo | | | | | 🟩 | 🟩 | 🟩 X | | |
| 버스 / bus | | | | | 🟩 | 🟩 X | 🟩 | | |
| 는 / neun | | | | | | | | 🟩 | 🟩 |
| 가 / ga | | | | | | | | 🟩 X | 🟩 |
| 시내 / sinae | | | | | | | | 🟩 | 🟩 X |

X Word alignment 🟩 Phrase alignment induced from word alignment □ Base phrase

(a) Word and phrase alignment

(b) Projecting a dependency tree

1. <"sinae ganeun" : "for downtown", "bus siganpyo" : "the bus timetable" : Reverse>
2. <"bus siganpyo" : "the bus time-table", "boyi eo_ju sigess seupnika" : "Would you show me" : Reverse>

(c) Extracting bilingual dependency relations

Fig. 1. Process of acquiring bilingual dependency relations.

sentence in the other language.

As input, a source language sentence is dependency parsed to the base phrase level, and a target language sentence is chunked by a shallow parser. First, for the word alignment, we use the GIZA++ toolkit [2]. For the phrase alignment, we utilize the word alignment and the base phrase boundary information in each language's sentence, keeping consistency between the word alignment and the phrase boundaries as in [4] (see Fig. 1(a)). Then, the bilingual dependency parse is acquired by sharing the dependency relations of a monolingual dependency parser among the aligned phrases. Finally, we extract bilingual dependency relations by traversing the dependency parse. Figures 1(b) and (c) show the process of bilingual parsing and the extraction of bilingual dependency relations between Korean and English. The extracted relation is a binary relation between a modifier and its head paired with their translation.

## 2. Acquiring Paraphrases Depending on Context

A basic concept of paraphrasing is the assumption that if multiple Korean phrases are equivalent to each other, they can be translated into a single English phrase. But, the reverse is not always true. That is, even though a single phrase maps onto multiple phrases in the foreign language, the phrases might not be paraphrases. This implies that the sense of the original phrase should be disambiguated depending on a given context.

For the extraction of paraphrases for which their meaning is disambiguated under a certain context, we try to give a strong constraint with bilingual context evidence of a dependency relation, denoted as R(x, y), where x is the head of a modifier y:

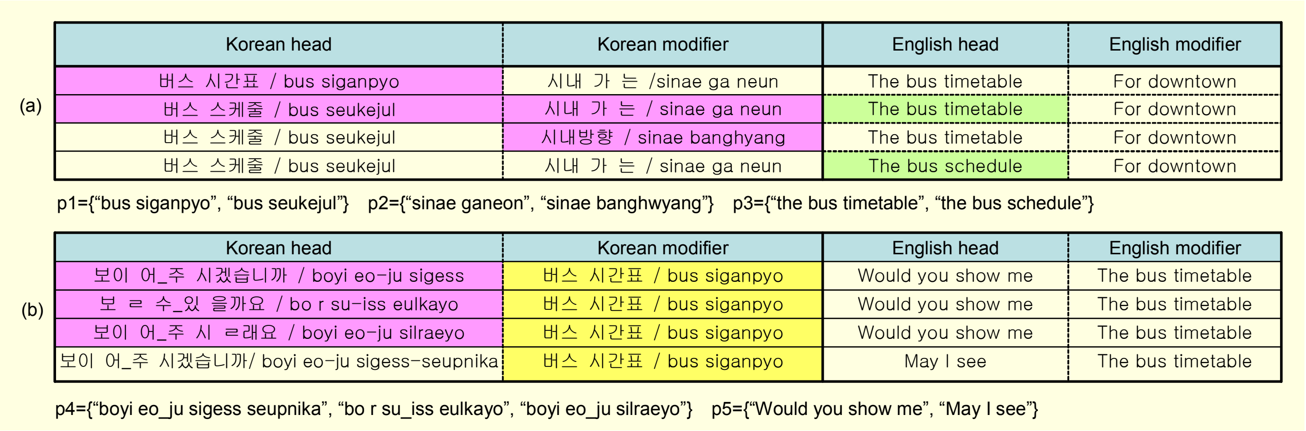
$$R(e_i,e_j) \Leftrightarrow R(k_{a_i},k_{a_j}) \wedge R(e_i,e_j) \Leftrightarrow R(k'_{a_i},k'_{a_j}) \\ \Rightarrow R(k_{a_i},k_{a_j}) \Leftrightarrow R(k'_{a_i},k'_{a_j}), \qquad (1)$$

$$R(e_i,e_j) \Leftrightarrow R(k_{a_i},k_{a_j}) \wedge R(k_{a_i},k_{a_j}) \Leftrightarrow R(k'_{a_i},k'_{a_j}) \\ \Rightarrow k_{a_i} = k'_{a_i} \quad \text{iff} \quad k_{a_j} = k'_{a_j}. \qquad (2)$$

For the identification of paraphrases, we equate the different dependency relations $R(k_{ai}, k_{aj})$ and $R(k'_{ai}, k'_{aj})$ aligned with a unique dependency relation $R(e_i, e_j)$ in the other language and regard them as a set of paraphrased dependency relations (see (1)). Under the constraint of the paraphrased dependency relations, we again try to acquire paraphrases at the phrase level. That is, we extract the phrases sharing the same head/modifier phrase in paraphrased dependency relations as a phrase paraphrase under a given bilingual dependency context (see (2)).

Figure 2 shows some examples of paraphrased dependency relations and phrase paraphrases. In Fig. 2(a), the Korean dependency relations <"bus siganpyo", "sinae ganeun">, <"bus seukejul", "sinae ganeun"> and <"bus seukejul", "sinae banghyang"> mapped onto the same English relation, <"the bus timetable", "for downtown"> might be the paraphrased relations. Under the condition of paraphrased dependency relations, the phrases, "bus seukejul" and "bus siganpyo" modified by the same modifier phrase "sinae ganeun" can be extracted as paraphrases.

The induced set of paraphrases can be applied to dependency relations to extend the set through higher inference as in Fig. 2(b). We replace a phrase, which is a part of a bilingual dependency relation and a member of a paraphrase set with the representative phrase of the paraphrase set. Then, we iteratively apply the paraphrase extraction algorithm to the bilingual dependency relations and can acquire new paraphrase sets, such as p4 and p5.



(a)

| Korean head | Korean modifier | English head | English modifier |
|---|---|---|---|
| 버스 시간표 / bus siganpyo | 시내 가 는 /sinae ga neun | The bus timetable | For downtown |
| 버스 스케줄 / bus seukejul | 시내 가 는 / sinae ga neun | The bus timetable | For downtown |
| 버스 스케줄 / bus seukejul | 시내방향 / sinae banghyang | The bus timetable | For downtown |
| 버스 스케줄 / bus seukejul | 시내 가 는 / sinae ga neun | The bus schedule | For downtown |

p1={“bus siganpyo”, “bus seukejul”} p2={“sinae ganeon”, “sinae banghwyang”} p3={“the bus timetable”, “the bus schedule”}

(b)

| Korean head | Korean modifier | English head | English modifier |
|---|---|---|---|
| 보이 어_주 시겠습니까 / boyi eo-ju sigess | 버스 시간표 / bus siganpyo | Would you show me | The bus timetable |
| 보 ㄹ 수_있 을까요 / bo r su-iss eulkayo | 버스 시간표 / bus siganpyo | Would you show me | The bus timetable |
| 보이 어_주 시 ㄹ래요 / boyi eo-ju silraeyo | 버스 시간표 / bus siganpyo | Would you show me | The bus timetable |
| 보이 어_주 시겠습니까/ boyi eo-ju sigess-seupnika | 버스 시간표 / bus siganpyo | May I see | The bus timetable |

p4={“boyi eo_ju sigess seupnika”, “bo r su_iss eulkayo”, “boyi eo_ju silraeyo”} p5={“Would you show me”, “May I see”}

Fig. 2. Paraphrase acquisition based on bilingual dependency relations.

## III. Experiments

We used the BTEC corpora, a collection of conversational travel phrases in Korean and English. We used 152,175 parallel sentences for training and 10,146 sentences for testing. Korean sentences were automatically dependency parsed by an in-house dependency parser, and English sentences were automatically chunked by an in-house shallow parser. Through alignment and bilingual dependency parsing, we extracted 66,664 bilingual dependency relations. As a result, 24.2% of the Korean phrases and 21.8% of the English phrases were paraphrased with more than two phrases with a given bilingual dependency context. To evaluate the accuracy of the acquired paraphrases, we randomly selected 100 sets of paraphrases for Korean and English phrases, respectively. Because the accuracy of paraphrases can vary depending on context, we selected the dependency relations that contained a phrase in a paraphrase set from the test set. Then we generated the dependency relations by substituting the phrase by the other paraphrases under the constraint of dependency context. Accuracy was judged by two native speakers for each language. We measured the percentage of completely interchangeable paraphrases. Table 1 shows the summary of experimental results.

Table 1. Experimental results.

| | Korean | English |
|---|---|---|
| No. of relations | 66,664 | |
| No. of unique relations | 59,633 | 58,187 |
| No. of unique phrases | 36,157 | 33,088 |
| No. of paraphrase sets | 6,156 | 5,390 |
| Paraphrasing ratio (%) | 24.2 | 21.8 |
| Accuracy (%) | **94.6** | **84.6** |
| Paraphrasing ratio (%) [3] | 44.4 | 37.4 |
| Accuracy (%) [3] | 71.4 | 76.2 |

For comparison, we evaluated the accuracy of a previous work [3], which does not use bilingual dependency context, on the same BTEC corpora. As shown in Table 2, our proposed method can acquire paraphrases of higher accuracy than [3], even with fewer extracted paraphrases.

## IV. Conclusion

We have proposed a method to extract paraphrases by utilizing bilingual dependency contexts, which are extracted through bilingual dependency parsing based on word and phrase alignments. Our method produces higher-quality paraphrases by clearly disambiguating the sense of original phrases. This method could drastically reduce the cost of expanding bilingual parallel corpora and acquiring translation knowledge.

## References

[1] P. Koehn, F.J. Och, and D. Marcu, “Statistical Phrase-Based Translation,” *Proc. of the Human Language Technology Conf. (HLT/NAACL)*, 2003, pp. 201-228.

[2] F.J. Och and H. Ney, “Improved Statistical Alignment Models,” *Proc. of the 38th Annual Meeting of the Association for Computational Linguistics*, 2000, pp. 440-447.

[3] C. Bannard and C. Callison-Burch. “Paraphrasing with Bilingual Parallel Corpora,” *Proc. of ACL*, 2005.

[4] Y.S. Hwang, A. Finch, and Y. Sasaki, “Improving Statistical Machine Translation Using Shallow Linguistic Knowledge,” *Computer Speech and Language*, vol. 21, no. 2, 2007.