Paraphrasing depending on Bilingual Context Toward Generalization of Translation Knowledge

Young-Sook Hwang
ETRI
161, Yuseong-gu, Daejeon
305-700, KOREA
yshwang7@etri.re.kr

YoungKil Kim
ETRI
161, Yuseong-gu, Daejeon
305-700, KOREA
kimyk@etri.re.kr

SangKyu Park
ETRI
161, Yuseong-gu, Daejeon
305-700, KOREA
parksk@etri.re.kr

Abstract

This study presents a method to automatically acquire paraphrases using bilingual corpora, which utilizes the bilingual dependency relations obtained by projecting a monolingual dependency parse onto the other language sentence based on statistical alignment techniques. Since the paraphrasing method is capable of clearly disambiguating the sense of an original phrase using the bilingual context of dependency relation, it would be possible to obtain interchangeable paraphrases under a given context. Also, we provide an advanced method to acquire generalized translation knowledge using the extracted paraphrases. We applied the method to acquire the generalized translation knowledge for Korean-English translation. Through experiments with parallel corpora of a Korean and English language pairs, we show that our paraphrasing method effectively extracts paraphrases with high precision, 94.3% and 84.6% respectively for Korean and English, and the translation knowledge extracted from the bilingual corpora could be generalized successfully using the paraphrases with the 12.5% compression ratio.

1 Introduction

Approaches based on bilingual corpora are promising for the automatic acquisition of translation knowledge. Phrase-based SMT(Statistical Machine Translation) models have advanced the state of the art in machine translation by expanding the basic unit of translation from words to phrases, which allows the local reordering of words and translation of multi-word expressions(Chiang, 2007) (Koehn et al., 2003) (Och and Ney, 2004).

However phrase-based SMT techniques suffer from data sparseness problems, that is; 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 for these problems is to utilize paraphrases. An unknown phrase can be replaced with its paraphrase that is already known. Moreover, we can smooth the phrase translation probability using the class of paraphrases.

On the other hand, EBMT or PBMT systems might translate a given sentence fast and robustly geared by sentence translation patterns or generalized transfer rules. Since it costs too much to construct the translation knowledge, they suffer from the problem of knowledge acquisition bottleneck.

In this study, we present a method of automatically extracting paraphrases from bilingual corpora. Furthermore, we introduce a new method for acquiring the generalized translation knowledge. The translation knowledge is a kind of verb subcategorization pattern composed of bilingual dependency relations. We obtain the generalized translation knowledge by grouping the equivalent constituent phrases. The task of identifying the phrases equivalent to each other is defined as paraphrasing.

Our paraphrasing method utilizes bilingual corpora and alignment techniques in SMT. Unlike pre-

vious approaches which identify paraphrases using a phrase in another language as a pivot without context information (Bannard et al., 2005), or apply the distributional hypothesis to paths in dependency trees for inferring paraphrasing rules from monolingual corpora(Lin et al., 2001), we take the bilingual context of a bilingual dependency relation into account for disambiguating the sense of paraphrases. First, we create a large inventory of bilingual dependency relations and equate the pairs of dependency relations that are aligned with a single dependency relation in the other language as paraphrased dependency relations. Then, we extract the phrases sharing the same head (or modifier) phrase among the paraphrased dependency relations aligned with a unique dependency relation in the other language. We regard them as conceptually equivalent paraphrases. This work is based on the assumption of similar meaning when multiple phrases map onto a single foreign language phrase that is the converse of the assumption made in the word sense disambiguation work(Diab and Resnik, 2002). 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.

In order to systematically acquire the generalized translation knowledge, our method includes following steps:

- Derive a bilingually parsed sentence through projecting the source language parse onto the word/phrase aligned target sentence.
- Extract bilingual dependency relations from the bilingual dependency parses.
- Acquire paraphrases by exploiting the extracted bilingual dependency relations.
- Generalize the bilingual dependency relations by substituting the phrases with their paraphrase class.

2 Extracting Translation Patterns

In this section, we introduce a method to acquire translation knowledge like a bilingual dependency pattern using bilingual corpus. The bilingual dependency pattern is defined as an asymmetric binary relationship between a phrase called head and another phrase called modifier which are paired with their corresponding translations in the other language. In order to acquire the bilingual dependency relations, we do bilingual dependency parsing based on the word/phrase alignments and extract bilingual dependency relations by navigating the dependency parse tree.

2.1 Bilingual Dependency Parsing based on Word/Phrase Alignment

Given an input sentence pair, a source language sentence is dependency parsed in a base phrase level and a target language sentence is chunked by a shallow parser. During the dependency parsing and the chunking, each sentence is also segmented into morphemes and we regard a morpheme as a word.

We make word alignments through the learning of IBM models by using the GIZA++ toolkit(Och and Ney, 2000): we learn the translation model toward IBM model 4, initiating translation iterations from IBM model 1 with intermediate HMM model iterations. For improving the word alignment, we use the word-classes that are trained from a monolingual corpus using the srilm toolkit(Stolcke, 2002). Then, we do phrase alignments based on the word alignments, which are consistent with the base phrase boundaries as well as the word alignments as (Hwang et al., 2007) did. A phrase is defined as a word sequence that is covered by a base phrase sequence, not by a single sub-tree in a syntactic parse tree.

After the word and the phrase alignments, we obtain bilingual dependency parses by sharing the dependency relations of a monolingual dependency parser among the aligned phrases. The bilingual dependency parsing is similar to the technique of bilingual parsing in a word level described in (Hwa et al., 2005)(Quirk et al., 2005). Our bilingual parsing in a phrase level has an advantage of being capable of reducing not only the parsing complexity but also the errors caused by structural differences between two languages, such like a Korean and English pairs[1].

For bilingual parsing between Korean and English, we use a Korean dependency parse on the

[1] Since we regard that a phrase in a source language sentence is aligned with a target phrase if at least one word in a source phrase is aligned with the words in a target phrase, we robustly project the source phrases onto the target phrases.

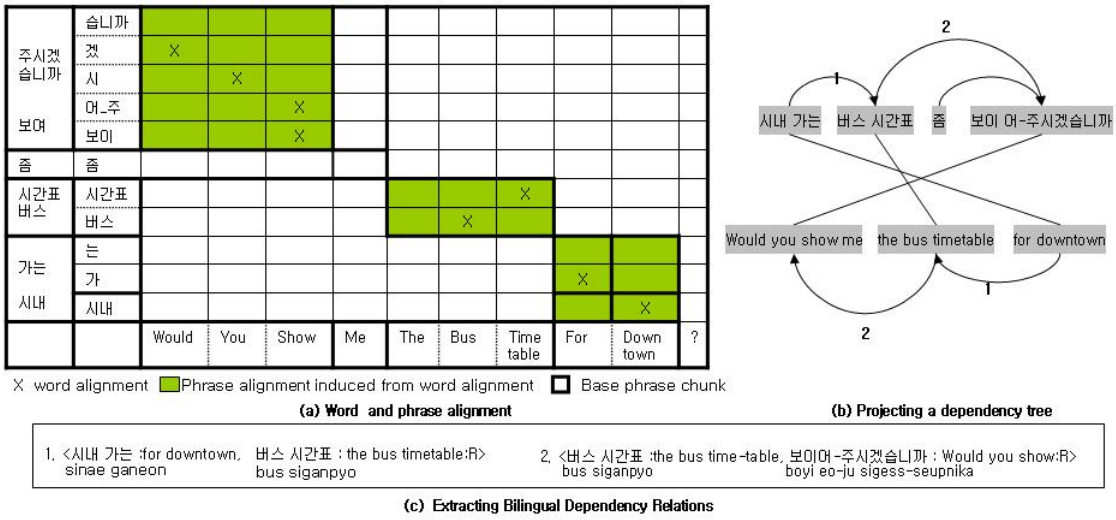


Figure 1: Illustration of Acquiring Bilingual Dependency Relations

source language side as a pivot. Figure 1 shows an illustration of bilingual dependency parsing between Korean and English based on the word/phrase alignments. The dependency structure induced on the target language side is in some sense isomorphic to the structure of the source language.

2.2 Extracting Bilingual Dependency Patterns

Starting from the head phrase of a given source language sentence, we extract bilingual dependency relations by traversing a bilingual dependency parse tree. A dependency relation is a binary relation between a head and modifier phrases. Each phrase is paired with its corresponding translation. For effectively using them during the decoding or the sentence generation, we attach an additional tag for indicating the order(e.g. Reverse or Forward) of target language phrases to the bilingual dependency relation. A dependency pattern refers to the bilingual dependency relation with the phrase order tag.

Figure 1(c) shows some examples of bilingual dependency patterns extracted from the bilingual dependency parse tree in Figure 1(b). In the example, Korean phrase ”sinae ga neun” aligned with the English phrase ”for downtown” modifies the phrase ”bus siganpyo” aligned with the English ”the bus timetable”. Through traversing the dependency parse trees, we acquire the bilingual dependency pattern <sinae ga neun:for downtown, bus siganpyo:the bus timetable;Reverse>.

If we apply the bilingual dependency pattern <sinae ga neun:for downtown, bus siganpyo:the bus timetable;Reverse> for machine translation of a given Korean expression ”sinae ga neun bus siganpyo”, we might generate an English phrase ”the bus timetable for downtown” by reversing the order of English head and modifier phrase corresponding to the Korean phrase ”sinae ga neun bus siganpyo”.

3 Acquisition of Paraphrases

Paraphrasing is based on 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 in a source language sentence maps onto multiple phrases in a foreign language sentence, the phrases might not be paraphrases. For example, two different Korean phrases, ”gyedan/{stairs,steps}” and ”baldongjak/steps”, might be translated into a single English phrase ”the steps”. But since the meaning of two Korean phrases is not equivalent to each other, the Korean phrases cannot be paraphrases. This implies that the sense of candidate paraphrases should be disambiguated depending on a given context.

For extracting the paraphrases of which sense is disambiguated under a given context, we give a strong constraint on paraphrases with bilingual context evidence of dependency relation denoted as *R(x, y)* :

Figure 2: Illustration of Paraphrasing based on Bilingual Dependency Relations

$$R(e_i, e_j) \equiv R(k_{a_i}, k_{a_j}) \text{ and } R(e_{i'}, e_{j'}) \equiv R(k_{a_{i'}}, k_{a_{j'}}) \Rightarrow R(k_{a_i}, k_{a_j}) \equiv R(k_{a_{i'}}, k_{a_{j'}}) \quad (1)$$

where the relation of $R(e_i, e_j) = R(e_{i'}, e_{j'})$ with the condition of $e_i = e_{i'}$ and $e_j = e_{j'}$.

$$R(e_i, e_j) \equiv R(k_{a_i}, k_{a_j}) \text{ and } R(k_{a_i}, k_{a_j}) \equiv R(k_{a_{i'}}, k_{a_{j'}}) \Rightarrow k_{a_i} \equiv k_{a_{i'}} \text{ iff } k_{a_j} \equiv k_{a_{j'}} \quad (2)$$

For the identification of paraphrases, we equate the different dependency relations aligned with a unique dependency relation in the other language and regard them as a set of paraphrased dependency relations (see eq.(1)). Under the constraint of the paraphrased dependency relations, we again try to acquire paraphrases at a 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 eq.(2)).

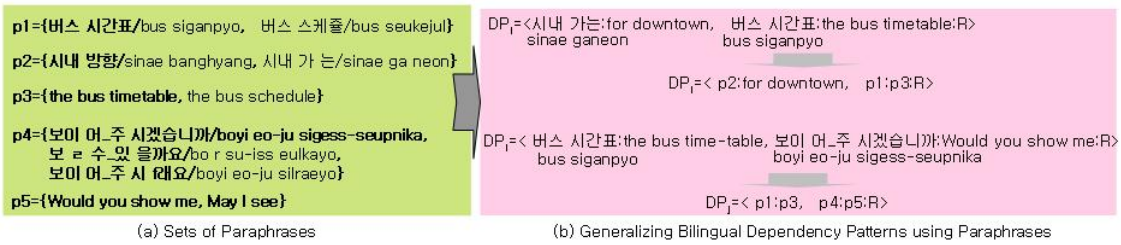
Figure 2 shows some examples of paraphrased dependency relations and paraphrases. In Figure 2 (a), the Korean dependency relations <bus siganpyo, sinae ga neun>,<bus seukejul, sinae ga neun> and <bus seukejul, sinae banghyang> mapped onto the English relation <the bus timetable, for downtown> are the paraphrases. Under the condition of paraphrased dependency relations, the phrases, "bus seukejul" and "bus siganpyo" modified by the same phrase "sinae ga neun" are extracted as paraphrases. In the same way, the set of modifier phrases, p1={"sinae banghyang", "sinae ga neun"} is acquired as a paraphrase set. For English, we obtain the set of paraphrases, p3={"the bus timetable", "the bus schedule"} as we did for Korean.

The induced set of paraphrases can be applied to dependency relations to extend the set through higher inference as in Figure 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. And we repeatedly apply the paraphrase extraction algorithm to the bilingual dependency relations of which a part is replaced with the previously acquired paraphrase set. Finally, we can acquire new paraphrase sets such as p4 and p5.

4 Generalizing Translation Patterns

The acquired paraphrases can be utilized for various NLP applications. In this work, we focus on making use of the paraphrases to generalize the translation knowledge of bilingual dependency patterns. By generalizing the bilingual dependency patterns, we aim at increasing the coverage of them without any over-generation.

The algorithm for generalizing bilingual dependency patterns is very simple. The main idea is to replace the constituent phrases of a given bilingual dependency pattern with their paraphrase classes. The paraphrase classes are extracted under the condition of a given bilingual context as follows: $< PP(k_m, dp_i) : PP(e_m, dp_i), PP(k_h, dp_i) : PP(e_h, dp_i); Order := Reverse|Forward >$ where the



(a) Sets of Paraphrases (b) Generalizing Bilingual Dependency Patterns using Paraphrases

Figure 3: Illustration of Generalizing Bilingual Dependency Patterns

function, $PP(x, y)$ returns the identifier of the paraphrase set of a given phrase x, which is constrained on a given context $y = dp_i$; k_m and k_h denote a modifier and a head in Korean, respectively and e_m and e_h denote the English phrases.

Figure 3 shows an illustration of generalizing the translation patterns using the previously acquired paraphrase classes. In the pattern dp_i, the English modifier "for downtown" uses the phrase itself because there is no paraphrase class. But, the others are generalized by using their paraphrase classes.

5 Experiments

We used the Basic Travel Expression Corpus (BTEC)(Takezawa et al., 2002), a collection of conversational travel phrases for Korean and English. We used 152,175 sentences in parallel corpus for training and 10,146 sentences for test. The Korean sentences were automatically dependency parsed by in-house dependency parser and the English sentences were chunked by in-house shallow parser.

Through experiments, we investigated the accuracy of the acquired paraphrases, and the compression ratio of the generalized translation patterns compared to the raw translation patterns. Moreover, we show the strength of utilizing bilingual context information in the acquisition of paraphrases with the comparison to the previous approach.

5.1 Accuracy of the Acquired Paraphrases

Through the alignments and bilingual dependency parsing, we extracted 66,664 bilingual dependency relations. 24.15% of Korean phrases and 21.8% of English phrases are paraphrased with more than two phrases under a given bilingual dependency context. The statistics of Korean and English paraphrases based on bilingual dependency relations is shown in Table 1.

Especially, the paraphrasing ratio of the Korean head phrases, 28.63% is higher than that of the English heads,22.6%. Many of the Korean head phrases are verb phrases that reflects the honorific and inflectional characteristics of Korean language. We might expect that the problems caused by various honorific expressions can be resolved with the paraphrases such like {"ga *r geoyeyo*", "ga *gess-seupnida*"}.

For evaluating the accuracy of the acquired paraphrases, we randomly selected 100 sets of paraphrases for Korean and English phrase respectively. Because the accuracy of paraphrases can vary depending on context, we selected the dependency relations that contain a phrase in a paraphrase set from the test set. And we generated the dependency relations by substituting the phrase by the other paraphrases. Accuracy was judged by two native speakers for each language. We measured the percentage of completely interchangeable paraphrases under a given bilingual dependency context.

Table 1 shows the performance of the paraphrases depending on their bilingual context. The accuracy of Korean and English paraphrases are 94.6% and 84.6% respectively. Korean paraphrases are more accurate than English paraphrases. Especially the quality of Korean head paraphrases(97.5%) is very high.

Since we used a simple base-phrase chunker for English, where most base phrases except for noun phrases are composed of single words, most of English phrases aligned to Korean phrases were dependent on the word alignments. Big structural difference between Korean and English made the word alignments more difficult. These alignment results might influence not only the paraphrasing ratio but

	Korean Relation		English Relation	
	Kor-head	Kor-mod	Eng-head	Eng-mod
# of relations	66,664		66,664	
# of uniq relations	59,633		58,187	
	36,157		33,088	
# of uniq phrases	17,867	22,699	13,623	24,000
	6,156		**5,390**	
# of paraphrase set	4,474	2,890	3,425	3,169
	24.15		**21.8**	
Paraphrasing Ratio(%)	28.63	17.7	22.6	19.4
	94.6		**84.6**	
Accuracy(%)	97.5	91.2	86	82.3
Paraphrasing ratio(%) (Bannard et al., 2005)	**44.4**		**37.4**	
accuracy (%) (Bannard et al., 2005)	**71.4**		**76.2**	

Table 1: Statistics of the extracted bilingual dependency relations and paraphrases

also the performance of the paraphrases.

Nevertheless, our paraphrasing method outperformed previous approaches which do not use bilingual dependency context. Because the paraphrasing methods are different, we could not compare them directly. But, we tried to make similar experimental condition on the same BTEC corpus by implementing the previous approach(Bannard et al., 2005). When evaluating the previous approach, the accuracy of (Bannard et al., 2005) was 71.4% and 76.2% for Korean and English paraphrases, respectively. The results show that our paraphrasing method can acquire the paraphrases of higher quality than (Bannard et al., 2005) while the paraphrasing ratio is lower than (Bannard et al., 2005).

5.2 Power of Generalization by Paraphrases

Finally, we investigated how many the extracted bilingual dependency patterns are generalized. Among 66,664 bilingual dependency patterns, 20,968 patterns were generalized into 12,631 unique generalized patterns by applying the extracted paraphrases[2]. As a result, the 66,664 bilingual dependency patterns were compressed into 58,324 generalized patterns with 12.5% compression ratio.

Furthermore, we examined how many bilingual dependency patterns can be generated by the generalized patterns in reverse. When replacing the generalized phrases with all of their paraphrases in both English and Korean sides, 235,640 bilingual translation patterns are generated. These are 3.53 times of the amount of the original translation patterns.

Even we have some errors in the paraphrase sets, these results might contribute to increasing the coverage of the translation knowledge for machine translation.

[2]A paraphrase set is composed of more than two paraphrases

6 Related Work and Discussion

The proposed paraphrasing method can be an extension of the work done by (Bannard et al., 2005). They introduced the method for extracting paraphrases: Using the automatic alignment method from phrase-based SMT, they showed that paraphrases in one language can be identified using a phrase in another language as a pivot. Furthermore, they defined a paraphrase probability to rank the extracted paraphrases and suggested a method to refine it by taking contextual information into account i.e. including simple language model.

Our study for paraphrasing is similar to their work but we take the bilingual dependency context into account for disambiguating the sense of a phrase. Limiting the candidate paraphrases to be the same sense as the original phrase is critical to the performance of paraphrases. Our approach provides the solution to clearly disambiguate the sense of a phrase using bilingual context information. This is the strong point of our approach different from the previous approaches.

Furthermore, in this work, we presented a method to acquire somewhat generalized machine translation knowledge of bilingual dependency patterns. There are few research of the acquisition of translation knowledge such like verb sub-categorization patterns (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. It takes such an advantage of generalized bilingual frame semantics. But, they have problems of appropriate mapping from lexical entries to word senses and obtaining correct example sentences.

In our approach to acquire the generalized bilingual translation patterns, a bilingual dependency pattern is one of the decomposed bilingual verb subcategorization patterns. It is possible to construct more complicated bilingual verb sub-categorization pattern by applying a kind of unification operation. In that case, we have the advantage of automatically disambiguating the word/phrase senses via the alignment techniques contrary to (Fung et al., 2004).

7 Conclusion

In this paper,we proposed a method to extract paraphrases using bilingual corpora, which utilizes the bilingual dependency relations obtained by projecting a monolingual dependency parse onto the other language sentence based on statistical alignment techniques. The advantage of our paraphrasing method is that it can produce paraphrases of high quality by clearly disambiguating the sense of an original phrase.

Furthermore, we suggested an advanced method to acquire generalized translation knowledge using the extracted paraphrases. With the bilingual dependency patterns generalized by the paraphrases, we aim at reducing the translation ambiguity, but also increasing the coverage of the translation knowledge. The experimental results showed that our generalization method is effective to achieve the goals.

In future, we will utilize the paraphrases based on bilingual dependency relations for increasing the amount of bilingual corpus and for smoothing the phrase probability table in statistical machine translation. Moreover, we plan to apply the acquired translation patterns, which are generalized by paraphrases, to various machine translation systems.

Acknowledgements

This work was supported by the IT R&D program of MIC/IITA, Domain Customization Machine Translation Technology Development for Korean, Chinese, and English.

References

Colin Bannard and Chris Callison Burch. 2005. *Paraphrasing with Bilingual Parallel Corpora*, Proc. of the 43rd Annual Meeting of the Association for Computational Linguistics.

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

David Chiang. 2007. *Hierarchical phrase-based translation*, Computational Linguistics, 33(2).

M. Diab and P. Resnik. *An Unsupervised Method for Word Sense Tagging Using Parallel Corpora*, Proc. of the 40th Annual Meeting of the Association for Computational Linguistics.

Atsushi Fujita, Kentaro Inui, and Yuji Matsumoto. 2005. *Exploiting Lexical Conceptual Structure for Paraphrase Generation*, Proc. of the 2nd International Joint Conference on Natural Language Processing (IJCNLP).

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

Rebeca Hwa, Philip Resnik, Amy Weinberg, Clara Cabezas and Okan Kolak. 2005. *Bootstrapping parsers via syntactic projection across parallel texts* , Natural Language Engineering , Vol 11(3), Pages: 311 - 325

Young-Sook Hwang, Andrew Finch and Yutaka Sasaki. 2007. *Improving statistical machine translation using shallow linguistic knowledge*, Computer Speech and Language , Vol. 21(2).

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003 *Statistical Phrase-Based Translation*, Proc. of the Human Language Technology Conference(HLT/NAACL)

D. Lin and P. Pantel 2001. *DIRT-Discovery of Inference Rules from Text*, ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pages 323-328.

Franz Josef Och and Hermann Ney. 2000. *Improved Statistical Alignment Models* , Proc. of the 38th Annual Meeting of the Association for Computational Linguistics, pp. 440-447, Hongkong, China.

Franz Josef Och and Hermann Ney. 2004. *The alignment template approach to statistical machine translation* , Computational Linguistics, Vol. 30(4), Pages 417-449.

C. Quirk, A. Menezes, and C. Cherry. 2005. *Dependency treelet translation: Syntactically informed phrasal SMT*, Proc. of the 43rd Annual Meeting of the Association for Computational Linguistics, pp. 271-279.

S. Stolcke 2002 *SRILM - an extensible language modeling toolkit*, Proc. of International Conference of Spoken Language Processing.

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