Grouping Synonymous Sentences from a Parallel Corpus

Hideki KASHIOKA

ATR Spoken Language Translation Research Laboratories
2-2-2 & Hikaridai “Keihanna Science City” Kyoto, Japan
hideki.kashioka@atr.jp

Abstract

Recently, natural language processing researches have focused on data or processing techniques for paraphrasing. Unfortunately, however, we have little data for paraphrasing. There are some research reports on collecting synonymous expressions with parallel corpus, though no suitable corpus for collecting a set of paraphrases is yet available. Therefore, we obtain a few variations of expression in paraphrase sets when we tried to apply this method with a parallel corpus. In this paper, we propose a grouping method based on the basic idea of grouping synonymous sentences related to the translation recursively, and decompose incorrect groups using the DM-decomposition algorithm. The incorrect groups include expressions that cannot be paraphrased because some words or expressions have different meanings in different situations. We discuss our method and experimental results with respect to BTEC, which is a multilingual parallel corpus.

1. Introduction

Synonymous expressions that convey the same information with different expression are useful for natural language understanding and/or natural language applications, such as summarization, machine translation, question answering, and so on. Therefore, resources of synonymous expressions are valuable. However, there are few resources of synonymous expressions, and it is difficult to determine whether many expressions are automatically grouped into synonymous groups, because many criteria can be defined from several viewpoints (lexically, syntactically, pragmatically etc.) for judging whether the expressions convey the same information. The relationship with translation pairs is one criterion for judging synonymous expressions. This criterion is effective for extracting synonymous sentence groups using sentence-aligned parallel corpus.

(Barzilay and McKeown, 2001) and (Shimohata and Sumita, 2002) mention the extraction of synonymous expression/sentence groups from a bilingual parallel corpus. They created synonymous groups in the one side of languages. In reality, few sentences in a synonymous group are grouped from their relationship with one source expression’s translation in a parallel corpus. Similarly to these synonymous sentence sets using one direction in relationships with translation pairs, we can obtain a group with another language part using another direction. When these synonymous groups are put together, we can find more expressions/sentences in a synonymous group. Of course, multilingual relationships can be expanded the groups, an expansion that brings about an increase of incorrect groups containing the multiple meaning. Therefore, we decompose the expanded groups based on the graph theory.

In this paper, we propose a method of grouping synonymous sentences in Section 2 and describe the grouping experiment with a multilingual parallel corpus in Section 3. We then discuss the method’s application in Section 4 and conclude this report. The corpus we use is the basic travel expression corpus (BTEC), which is constructed by ATR (T.Takezawa et al., 2002). BTEC is a sentence-aligned multilingual parallel corpus. In this study, we use it on 4 languages (Japanese, English, Korean, Chinese).

2. How to extract a synonymous sentence group

This section describes the method of grouping synonymous sentences using their relationships with the translation pairs (concatenation step) and the method for decomposing the set of synonymous sentences (decomposition step).

2.1. Concatenation step

The basic idea is very simple for concatenating expressions into groups. When the expression Exp_1^A written by language A and Exp_1^A is translated into the expressions, Exp_1^B, Exp_2^B, ...,Exp_n^B, by language B, the set of expressions Exp_1^B, Exp_2^B, ...,Exp_n^B make one synonymous group. Furthermore, when the sentence Exp_2^A written by language A and Exp_2^A is translated into the sentences Exp_1^B, Exp_{n+1}^B, ...,Exp_m^B by language B, the set of sentences Exp_1^B, Exp_{n+1}^B, ...,Exp_m^B ($n < m$) make one synonymous group. In this situation, Exp_1^A and Exp_2^A make a synonymous group because both Exp_1^A and Exp_2^A has a relationship with the translation pair of Exp_1^B. Thus, Exp_1^A and Exp_2^A in language A, and Exp_1^B, ...,Exp_m^B in language B make the synonymous group. If other language information is available, we can extend this synonymous group using information about translation pairs for other language.

Through the concatenation step, it is possible to display the relationship between two languages in the synonymous groups on a bipartite graph by graph theory, where each node indicates a sentence and a link indicates a translation pair, as in Figure 1.

2.2. Decomposition step

In some cases, there are incorrect groups constructed by the concatenation step due to words or phrases with multiple meanings.Then, we can decompose this graph with a Dulmage-Mendelsohn decomposition. The DM decomposition decomposes the graph into several components, and each component of each sub-group from the DM decomposition is strongly connected (except the two sub-groups placed first and last in a partially ordered graph). In other

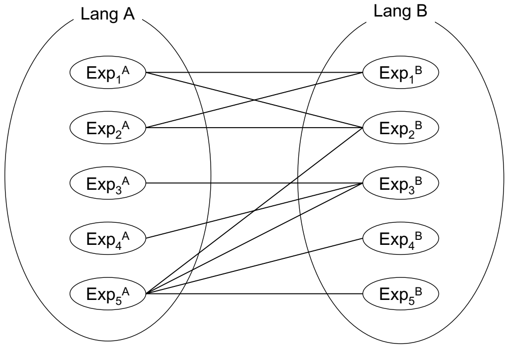

Figure 1: Constructed group

words, any two sentences in the same sub-group have more than two relationships with translation paths.

The graph in Figure 1 is decomposed into two graphs with DM decomposition, as in Figure 2. One of the graphs includes the four expressions Exp_1^A, Exp_2^A, Exp_1^B, Exp_2^B.

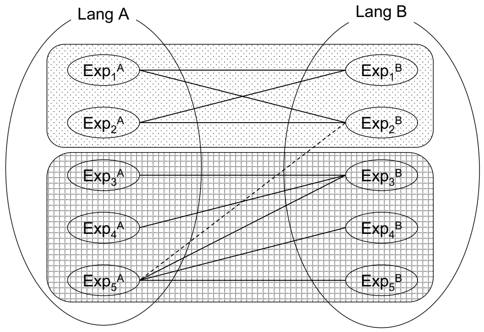

Figure 2: Decomposition group

Two expressions in this sub-group, Exp_1^A , Exp_1^B have two relationships: one is a relationship with a directory translation path, and the nother is a relationship like $Exp_1^A \rightarrow Exp_2^B \rightarrow Exp_2^A \rightarrow Exp_1^B$. Any other pair of two expressions in this sub-group can be found on two paths. The other graph includes six expressions, but the graph of these expressions has no strong connections.

3. Characteristics of an extracted synonymous sentence groups

This section explains the target corpus and shows the characteristics of the groups in proposed method.

3.1. Characteristics of BTEC

In this paper, we use a multilingual parallel corpus called the BTEC. This parallel corpus is a collection of Japanese sentences and their English, Korean, Chinese translations that are often found in phrase books for foreign tourists[1]. These sentence pairs cover a number of situations (e.g., hotel reservations, troubleshooting) for Japanese going abroad. In this paper, we use a part of the BTEC about 162,318 sentence pairs. Our using corpus contains 93,475 different Japanese sentences, 86,231 different English sentences, 94,382 different Korean sentences, and 82,171 different Chinese sentences. In the BTEC, some Japanese/English/Korean/Chinese sentences are exactly the same expressions as other pairs, whereas there are also pairs that consist of exactly the same Japanese sentences that are not always the same as English sentences.

3.2. Statistics groups

There are 69,255 groups with Japanese-English pairs and 61,463 groups using all four languages in the BTEC after the concatenation step. Figure 3 shows one of the groups made after the concatenation step using Japanese and English parallel data.

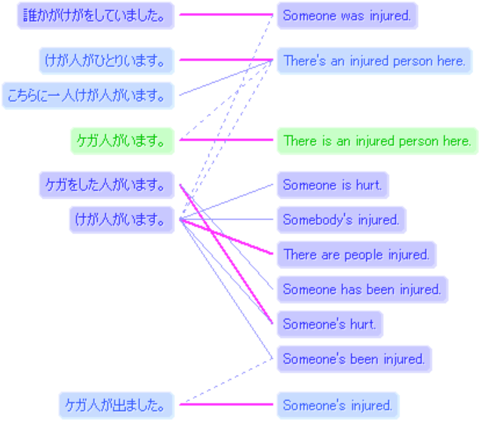

Figure 3: Example for Concatenation group

Quantity of groups

From the perspective of Japanese sentences, the following Table 1 shows a part of the distribution of the groups. These groups include a group that has only one translation pair because the BTEC includes sentences that do not appeare twice or more in the BTEC.

Number of Japanese sentences	Number of groups (language set)			
	(JE)	(JK)	(JC)	(JEKC)
1	62,427	76,534	75,195	54,291
2	4,244	4,077	4,378	4,277
3	1,152	946	1,177	1,187
4	457	356	476	515
5	246	190	230	288
6	144	98	148	173
7	123	73	64	142
8	65	42	42	101
9	42	35	30	63
≥ 10	355	109	90	426
total	69,255	82,460	81,730	61,463

Table 1: Distribution of the number of groups with the Number of Japanese sentences

Then, 7,172 really synonymous groups (that include more than two sentences in one language) can be obtained when using all four languages in the BTEC after the concatenation step. The groups using the JE set make more

[1] Currently ATR is extending the corpus to other languages.

groups that include many more sentences per group than do JK or JC sets;in fact, JK and JC sets are outnumbered by JE in the groups that include one sentence. When we extended the language set to concatenate the group, the number of groups decreases. The number of groups that include one sentence mainly decreases.

Figure 4 shows the relationship between the number of groups and the expression in the group for each language. Each language indicates similar relations between the number of groups and the sentences.

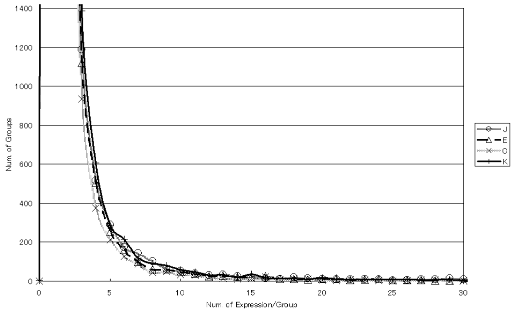

Figure 4: Relationship between the number of groups and the expression in the group for each language in all four languages used for grouping

Figure 5 shows the distribution of the number of groups dependent on the number of sentences for each language, where X means the number of Japanese sentences in the group, Y means the number of English sentences, and Z means the number of groups. This distribution is calculated from the groups by using all four languages. Figure 5 also shows the relation between Japanese and English. The relations between other language pairs are similar to the distribution of the relation between Japanese and English.

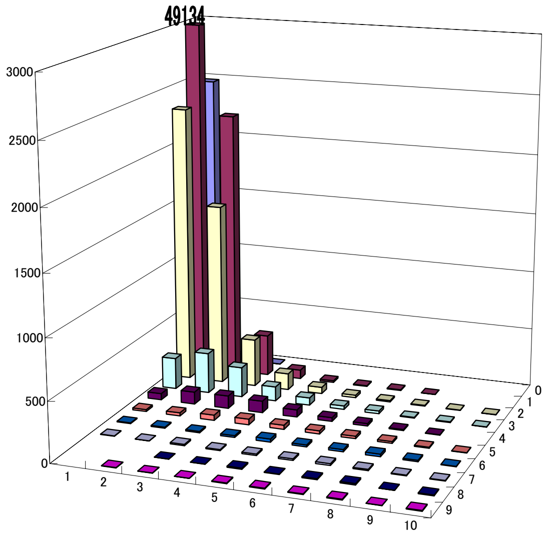

Figure 5: Distribution of the number of sentences between two languages

Quality of groups

The groups formed by the concatenation step are not completely synonymous groups; some of them include sentences with different meanings. We checked whether the grouped Japanese sentences are synonymous manually. We judge that the group is correct when all sentences have almost the same meaning and can be used in the same situation, and any others are wrong. We checked the part of groups (2,302 groups), that were selected from the groups made using Japanese and English, containing more than three Japanese sentences selected from parallel data. We then found that all but 72 groups contained almost synonymous sentences. These other 72 groups comprised three types of situation: first case contained words that have different meanings to the situation in the group, i.e. “It's hot, ” where some sentences in the group mean “It's spicy” and some mean “Hight temperature.” The second case contained sentences that had mostly the same meaning, though some sentences mentioned detailed situations, i.e. mainly in the group for requesting recommendations, some sentences included the expression meaning “with menu for dinner,” and some meaning “with tour for travelling.” The third case is included sentences with totally different meanings that were connected to very simple or ambiguous expressions, i.e. in the “When?” group. This linked to some sentences meaning “When do we start boarding?”, some meaning “What time does the store open?”. However, these incorrect groups can be partitioned into sub-groups by the decomposition step. Figure 6 shows a group that can be decomposed into three sub-groups.

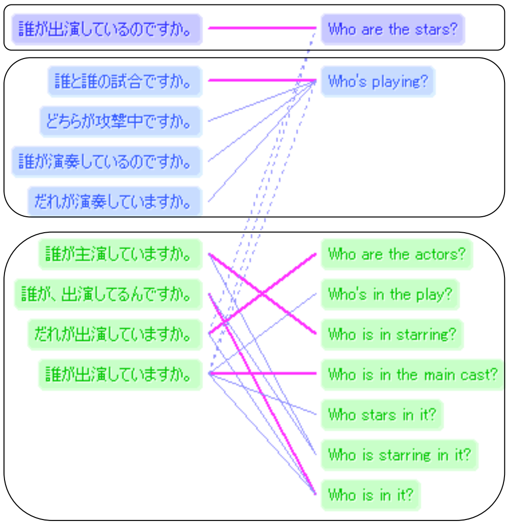

Figure 6: Example of Decomposed group

Decomposed steps are separated into approximately two sentences on average, with a maximum of 38 sentences per sub-group. All 72 groups could be properly partitioned when checked manually.

Structure of groups depends on languages used

Each synonymous group can be shown on a graph where a node is a sentence and a link between two nodes is a translation pair. Figure 7 displays the image of a group with concatenation steps using three languages. Forcing on language B of the synonymous groups in Figure 7, there are two groups concatenated with language A, and three groups with language C.

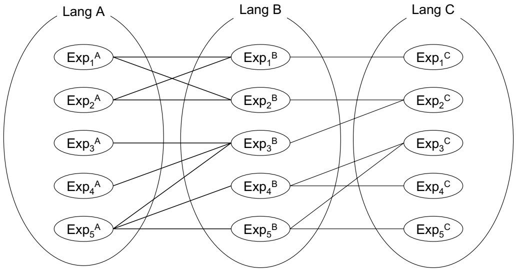

Figure 7: Image of the group by concatenation steps for three languages

Exp_2^B and Exp_3^B are important expressions because these expressions are overlapped in the groups concatenated with different languages.

4. Discussion

The proposed method enables us to obtain synonymous sentence groups. In the concatenation step, sentences with the same expression in the corpus were assigned to one node on the graph, as in Figure 8. Forcusing on the node, we can assign different expressions to one node on the graph if those different expressions are synonymous expressions. Therefore, we have the chance to extend its application to another corpus that does not include as many exactly matching expressions.

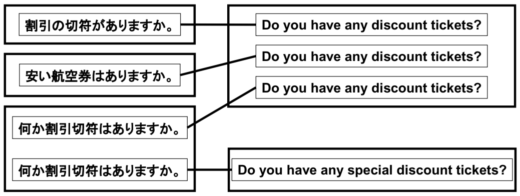
割引の切符がありますか。
安い航空券はありますか。
何か割引切符はありますか。
何か割引切符はありますか。
Do you have any discount tickets?
Do you have any discount tickets?
Do you have any discount tickets?
Do you have any special discount tickets?

Figure 8: Detailed concatenation graph

These synonymous sentence groups can be used as evaluation sets for a machine translation system. Recently, calculation for the BLEU score(Papineni et al., 2001), which is MT evaluation method, required several target sentences for test data. These synonymous groups are appropriate test sets.

From a different point of view, each group indicates one meaning on the translation, and each sentence in BTEC are assigned a situation label. When these groups are used for an example-based machine translation system, the output sentence selects suitable sentences from sentences in the most appropriate group for the situation.

From another point of view, syntactical or lexical synonymous information can be extracted with these synonymous groups.

5. Conclusion

In this paper, we proposed a method of grouping synonymous sentences with a multilingual parallel corpus, and displayed the distribution of the number of the sentences in each group. The grouping method is constructed with two steps: the first is a concatenation step using translation pair information, and the second is a decomposition step based on the graph theory. The BTEC is grouped into about 60,000 synonymous sentence groups, and almost all groups contain sentences with similar meanings. In the case of including sentences with different meanings in a group, the decomposition step forms a partition between sub-groups with different meanings. Therefore, these groups are useful for MT evaluation or development, and effective for establishing paraphrase rules.

6. Acknowledgement

The research reported here was supported in part by a contract with the Telecommunications Advancement Organization of Japan entitled, “A study of speech dialogue translation technology based on a large corpus.”

7. References

Barzilay, Regina and Kathleen R. McKeown, 2001. Extracting paraphrases from a parallel corpus. In ACL (ed.), *Proceedings of ACL-EACL’01*.

Papineni, K., S. Roukos, T. Ward, and W. Zhu, 2001. Bleu: a method for automatic evaluation of machine translation. Technical report.

Shimohata, Mitsuo and Eiichiro Sumita, 2002. Automatic paraphrasing based on parallel corpus for normalization. In *LREC-2002*.

T.Takezawa, E.Sumita, F.Sugaya, H.Yamamoto, and S.Yamamoto, 2002. Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. In *LREC-2002*.