A Method for Distinguishing Exceptional and General Examples in Example-based Transfer Systems

Hideo Watanabe
IBM Research, Tokyo Research Laboratory
1623-14, Shimotsuruma, Yamato-shi, Kanagawa-ken 242, JAPAN
e-mail: watanabe@trl.vnet.ibm.com

Abstract

Distinguishing exceptional translation examples is an important issue in example-based transfer systems, because such systems use exceptional and general translation examples uniformly. This paper describes a mechanism for dealing with exceptional translation examples in our example-based transfer system, *SimTran*, and proposes a method for identifying such examples in a translation example-base.

1 Introduction

In recent years, the example-based approach has been used in many areas of natural language processing [3, 7, 8, 10, 9, 1]. We have been using this approach to develop a transfer system called *SimTran* [13, 14, 16]. However, a bottleneck occured in the collection of large numbers of translation examples consisting of pairs of parsed structures in the source and target languages (hereafter we call these structures *translation patterns*), because parsing is not a perfect process. We now have some methods for overcoming this problem. For instance, recent studies [2, 11, 6, 12] have proposed mechanisms for collecting pairs of parsed structures automatically from translation examples, and in the previous paper [15], I proposed a method for extracting relevant translation patterns by comparing a wrong translation result and its correct translation. Using these methods, we can now collect translation patterns relatively easily.

There is, however, another problem called *example interference*, which means that an exceptional (or idiomatic) translation pattern is selected when a general translation pattern should be selected; this has a side-effect on the construction of a target structure. Suppose that we have the following two translation examples from Japanese to English (e1) and (e2),

(e1) watashi(I) ha konpyuutaa(computer) wo kyouyousuru.
I share the use of a computer.

(e2) watashi(I) ha kuruma(car) wo tsukau.
I use a car.

and that we are given the following Japanese input sentence (s1):

(s1) watashi(I) ha dentaku(calculator) wo shiyousuru.

In the above examples, (s1) is likely to be more similar to (e1) than (e2), because the three Japanese verbs "kyouyousuru," "tsukau," and "shiyousuru" are all very similar,[1] and "dentaku" ("calculator") is more similar to "konpyuutaa" ("computer") than "kuruma" ("car"). If this is the case, the English output obtained by using (e1) is (t1),[2] whereas it should be (t2):

(t1) I use the use of a calculator.

(t2) I use a calculator.

This problem occurs because example-based transfer systems choose examples simply on the basis of similarity. This can be considered by using the analogy of cells like those shown in Figure 1. In the figure, a dot represents a translation example, and a cell represents a space in which an input is determined to be similar. According to this analogy, an example-based system checks the cell in which an input is located, and uses an example governing the cell. If a new example is added in this space, a cell for it is created as if cell division. If an input happens to fall into the cell of an exceptional example, it is wrongly translated. Therefore, an exceptional example should be added as a special cell (a shaded dot in Figure 1) that has no extent in the example-based space, so that it cannot be used unless it matches the input exactly. Thus, an example-based transfer system must deal with exceptional translation patterns separately when calculating similarity.

This paper describes a mechanism used in *SimTran* for dealing with exceptional translation patterns in the same framework as general translation patterns, and proposes a method for identifying exceptional translation patterns in a translation pattern base.

The next section describes a mechanism for dealing with such translation patterns, and Section 3 de-

[1] Actually, they are in the same category (or the same leaf) in the Japanese thesaurus Bunrui-Goi-Hyou [5].

[2] The main verb is changed from "share" to "use," because "share" is not a translation of "shiyousuru."

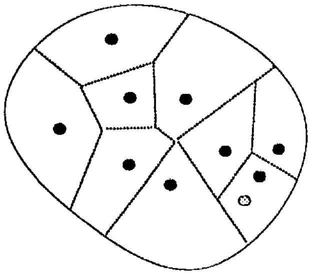

Figure 1: An example-base space

scribes a method for identifying exceptional translation patterns. Some experiments are reported in Section 4, and some issues are discussed in Section 5. Finally, some concluding remarks bring this paper to an end.

2 Mechanism for dealing with exceptional translation patterns

SimTran calculates the similarity between a subgraph of an input structure and the source part of a translation pattern on the basis of both the structural similarity and the similarity of the lexical-forms of corresponding nodes. For instance, the distance (the inverse of similarity) between two Japanese lexical-forms is expressed by the difference of their values in a Japanese thesaurus called Bunrui-Goi-Hyou [5][3] as follows:

$$distance(w_1, w_2) = \frac{|bghcode(w_1) - bghcode(w_2)| + \delta}{bghmax + \delta}$$

where $bghcode(w)$ is the code value in the Bunrui-Goi-Hyou, $bghmax$ is the maximal difference of the bghcodes, and δ is a penalty value incurred when w_1 and w_2 are not identical. This equation is used for lexical-forms in general translation patterns. If one is a lexical-form which requires exact-match in an exceptional translation pattern, then the distance is calculated as follows:

$$distance(w_1, w_2) = \begin{cases} 0 & w_1 \textit{ is identical to } w_2 \\ 1 & \textit{otherwise} \end{cases}$$

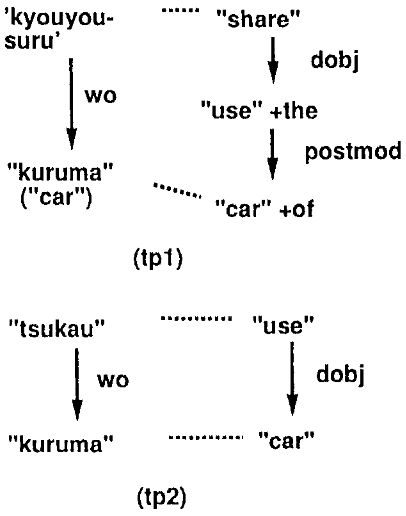

Figure 2: Exceptional translation pattern and general translation pattern

A lexical-form has a distinctive feature that makes it possible to determine which equation should be used in calculating similarity; if one of two lexical-forms is expressed by a single-quoted string, then the distance between the lexical-forms is calculated by using the second equation; on the other hand, if both lexical-forms are expressed by double-quoted strings, then their distance is calculated by using the first equation.

Thus, an exceptional translation pattern is distinguished by having nodes whose lexical-forms are single-quoted strings in its source part, while a general translation pattern is distinguished by having nodes whose lexical-forms are all double-quoted strings in its source part. Not all nodes in the source part of an exceptional translation pattern are necessarily single-quoted strings; single-quoted string nodes and double-quoted string nodes may be mixed in a translation pattern. In Figure 2, (tpl) is an exceptional translation pattern and (tp2) is a general translation pattern. Note that the root node of the Japanese part is the only single-quoted string in (tpl), and it matches only an input whose root node is 'kyouyousuru.'

By using this distinction of lexical-forms, we can integrate exceptionality handling into the similarity calculation framework without separating this task as a pre-process or post-process.

[3] Bunrui-Goi-Hyou is a Japanese thesaurus consisting of large trees for nominals, adjectives, and verbs. Each node is assigned a unique number. Similar concept words are located in similar positions (or assigned similar numbers) in these trees.

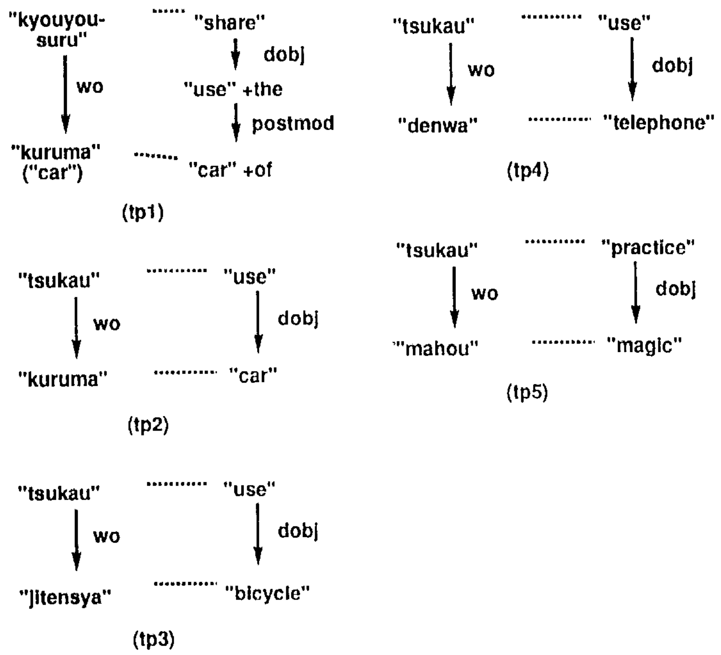

Figure 3: Example of the identification of exceptional translation patterns

3 Method for identifying exceptional translation patterns

For most people, an exceptional translation pattern is likely to mean a pattern of translation for an idiomatic or colloquial expression. In general, an idiomatic translation pattern is a translation pattern whose target part is markedly different from that of translation patterns whose source parts are similar to that of the idiomatic pattern. From the viewpoint of the transfer process, what we would like to identify are translation patterns that may have side-effects when they are selected instead of general translation patterns. We call such translation patterns *exceptional translation patterns*. According to this definition, exceptional translation patterns are not restricted to idiomatic patterns, in fact, more translation patterns other than idiomatic ones fall into this category. Here, we classify exceptional translation patterns into the following two categories:

- Extra-Exceptional Translation Patterns: These have some extra elements in the target part in addition to those in similar translation patterns.
- Intra-Exceptional Translation Patterns: These are almost same as similar translation patterns, but several target words are different.

When exceptional translation patterns are found, it is important to know whether two translation patterns are equivalent or not. Therefore, *equivalent* translation patterns are defined as follows:

Given two dependency structures d_1 and d_2, then they are called *equivalent* if and only if they are structurally identical and corresponding nodes have the similar semantic code.[4] Further, given two translation patterns $tp_1 = \langle s_1, t_1, m_1 \rangle$ $tp_2 = \langle s_2, t_2, m_2 \rangle$, where s_i is a source part, t_i is a target part, and m_i is a mapping from s_i to t_i, then these two translation patterns are called *equivalent* if they satisfy the following conditions:

(1) Both source parts are equivalent, and both target parts are structurally identical.

(2) The roots of t_1 and t_2 are the same string.

(3) For each node n in s_1, $m_1(n)$ is one of translation words of n.

(4) For each node n in s_2, $m_2(n)$ is one of translation words of n.

The algorithm for identifying exceptional translation patterns is as follows:

[4] For instance, the semantic code in Japanese is Bunrui-Goi-Hyou code. The extent to which two words are determined to be similar is also a parameter. It may vary according to the system. In this paper, two words are determined to be similar if they have the same semantic code.

Step 1 Divide translation patterns into several groups, each of which consists of equivalent translation patterns.

Step 2 For each pair of distinct translation pattern groups g_1 and g_2, if any pattern of g_1 is equivalent to any pattern of g_2 other than nodes governed by the root of the source part, then the translation patterns in g_1 and g_2 are marked *general.*

Step 3 For each pair of distinct translation pattern groups g_1 and g_2, if the source part of any pattern (p_1) of g_1 is equivalent to the source part of any pattern of g_2, but target parts of them are not structurally identical, because p_1 has extra elements, then the translation patterns of g_1 are marked *extra-exceptional.*

Step 4 For each non-exceptional translation pattern group g_1, if there is another general translation pattern group g_2 such that any pattern (p_1) of g_1 is equivalent to any pattern of g_2 other than the root node in the target part of p_1, then the translation patterns of g_1 are marked *intra-exceptional.*

Step 2 identifies possible general translation patterns if they are used in a relatively wide range of words, because in general an exceptional pattern is restricted in the usage of words. This approach, however, is not perfect for identifying general translation patterns, because there is a case such that the exceptionality derives from a single special word. Therefore, in the next step, checking does not exclude these possible general translation patterns. Step 3 identifies extra-exceptional translation patterns by checking the structure of the target part. Step 4 then identifies intra-exceptional ones by comparing the root node in the target part with the root nodes in the target part of possible general translation patterns. The reason why this comparison is restricted to possible general translation patterns is that intra-exceptional translation patterns have side-effects only when they are similar to general translation patterns.

Figure 3 shows an example of the identification of exceptional translation patterns, in which the Japanese verbs "kyouyousuru" and "tsukau" have the same bghcode, and the Japanese nouns "kuruma," "denwa," and "mahou" have different bghcodes, on the other hand, "kuruma" and "jitensya" have the same bghcode. First, step 1 divides these translation patterns into four groups: group 1 consists of (tp1), group 2 consists of (tp2) and (tp3), group 3 consists of (tp4), and group 4 consists of (tp5). Step 2 identifies group 2 and 3 as general translation patterns, because "kuruma" and "denwa" have different bghcodes. Subsequently, step 3 identifies (tp1) as an extra-exceptional translation pattern, because (tp1) has extra elements "the use of" for (tp2). Further, step 4 identifies (tp5) as an intra-exceptional translation pattern, because (tp5) is equivalent to the general translation patterns (tp2), (tp3) and (tp4), other than "use" and "practice" in the root nodes of the target parts.

4 Experiments

We have tested the above-mentioned algorithm with translation patterns in a Japanese-to-English transfer dictionary that was previously used in our laboratory. For each bghcode, we collected translation patterns such that the root of the source part has the code, and applied the algorithm to the translation pattern set of each category. Table 1 shows the resulting top 10 categories with respect to the total number of occurrences. In most categories, more than 90% of translation patterns were identified as exceptional. The reason for the lopsidedness of this result is that the translation patterns described in the previous transfer dictionary were almost all exceptional cases that could not be dealt with by the default procedures coded in the transfer module. Therefore, this result indicates that the algorithm is able to idenitfy exceptional translation patterns correctly.

5 Discussion

In conventional transfer systems [4], transfer rules are roughly divided into general ones and exceptional (or idiomatic) ones. The transfer system checks the exceptional cases first, and if they cannot match the input then the system applies general rules. On the other hand, example-based transfer systems deal with translation patterns (or examples) uniformly on the basis of similarity, according to the example-based principle. This mechanism causes the example interference problem. A very useful property of the example-based approach is that it allows a sentence to be added as an example if it cannot be dealt with properly. This holds if the same input as the newly added example is given, but when the resolution of the similarity calculation is not enough, an input that is similar to but not exactly the same as the added example may not be dealt with properly, because there may be another similar example that is exceptional. Therefore, it is very important to identify whether an example is general or exceptional.

After application of the algorithm described in this paper, translation patterns are classified into the following categories: general, exceptional (extra- and intra-), and neutral. Neutral translation patterns, which are not marked general or exceptional, are

Bghcode (example)	Num of Total	Num of General	Num of Exceptional (extra, intra)	Exceptional (extra only) /Total
15210(idousuru)	247	1	232 (228, 4)	93% (92%)
15270(iku)	174	0	138 (137, 1)	79% (78%)
15310(torikomu)	165	0	160 (150, 10)	96% (90%)
15600(tikazuku)	199	1	185 (178, 7)	92% (89%)
15710(kiru)	185	0	181 (159, 22)	97% (85%)
30110(kurushimu)	192	8	183 (160, 23)	95% (83%)
30200(suki)	280	6	271 (203, 68)	96% (72%)
30610(omou)	180	0	179 (169, 10)	99% (93%)
31200(iu)	191	0	173 (173, 0)	90% (90%)
36700(hattyuu)	182	0	181 (168, 13)	99% (92%)
38520(tsukau)	65	2	60 (53, 7)	92% (81%)

Table 1: Experimental results for transfer dictionary

translation patterns that do not have side-effects. They are not used for a wide variety of words in the current translation pattern base. If more translation patterns are added later, they may be identified as general or exceptional. By this method, one can enable the system to identify exceptional translation patterns automatically by adding some general translation patterns similar to them. This is a very useful feature for bootstrapping of a translation pattern base. A weak point of this algorithm, however, is that it requires a large number of translation patterns. If enough translation patterns are not given, exceptional translation patterns might not be identified. However, collecting many translation patterns is no longer a serious problem, since several methods for collecting them automatically have been proposed in recent studies [2, 11, 14, 6].

The method proposed in this paper probably does not comply with human intuition regarding idiomatic translation patterns; rather, it detects translation patterns that are idiomatic for the system, in other words, patterns that might have side-effects in the current set of translation patterns. It probably requires deeper semantic processing to identify translation patterns that are idiomatic in the conventional sense.

6 Conclusion

In this paper, we have shown a problem of example-based transfer systems, *example interference*, and described a mechanism for dealing with exceptional translation patterns and general translation patterns uniformly in similarity calculation without destroying the whole framework of example-based processing. Further, we have proposed a method for distinguishing exceptional translation patterns from general translation patterns. In some cases, this method gives results that do not match human intuition regarding idiomatic translation patterns, but it can detect, from the viewpoint of example-based processing, translation patterns in the current translation pattern base that might have side-effects.

References

[1] Furuse, O. and Iida, H., "Cooperation between Transfer and Analysis in Example-Based Framework," *Proc. of Coling 92,* Vol. 2, pp. 645–651, 1992.

[2] Kaji, H., Kida, Y., and Morimoto, Y., "Learning Translation Templates from Bilingual Text," *Proc. of Coling 92,* Vol. 2, pp. 672–678, 1992

[3] Nagao, M., "A Framework of a Mechanical Translation between Japanese and English by Analogy Principle," Elithorn, A. and Banerji, R. (eds.) : *Artificial and Human Intelligence,* NATO 1984.

[4] Nagao, M., "The Transfer Phase of the Mu Machine Translation System," *Proc. of Coling '86,* pp. 97–103, 1986.

[5] National Language Research Institute: Bunrui Goi Hyou (in Japanese), Syuuei Syuppan, 1964.

[6] Matsumoto, Y., Ishimoto, H., and Utsuro, T., "Structural Matching of Parallel Text," *Proc. of 31st Annual Meeting of ACL,* pp. 23–30, 1993.

[7] Sadler, V., "Working with Analogical Semantics," Foris Publications, 1989.

[8] Sato, S. and Nagao, M., "Toward Memory-based Translation," *Coling 90,* 1990.

[9] Sato, S., "Memory-based Translation," *Doctor Thesis,* 1992.

[10] Sumita, E., Iida, H., and Kohyama, H., "Translating with Examples: A New Approach to Machine Translation," *Proc. of Info Japan 90,* 1990.

[11] Utsuro, T., Matsumoto, Y., and Nagao, M., "Lexical Knowledge Acquisition from Bilingual Corpora," Proc. of Coling '92, Vol. 2, pp. 581–587, 1992.

[12] Utsuro, T., Matsumoto, Y., and Nagao, M., "Verbal Case Frame Acquisition from Bilingual Corpora," Proc. of IJCAI '93, Vol. 2, pp. 1150–1156, 1993.

[13] Watanabe, H., "A Model of a Transfer Process Using Combinations of Translation Rules," *Proc. of Pacific Rim of Int. Conf. on AI '90,* 1990.

[14] Watanabe, H., "A Similarity-Driven Transfer System," Proc. of Coling '92, Vol. 2, pp. 770–776, 1992.

[15] Watanabe, H., "A Method for Extracting Translation Patterns from Translation Examples," *Proc. of 5th Int. Conf. on Theoretical and Methodological Issues in Machine Translation,* pp. 292–301, 1993.

[16] Watanabe, H. and Maruyama, H., "A Transfer System Using Example-Based Approach," IEICE Transactions on Information and Systems, Vol. E77-D, No. 2, pp. 247–257, Feb. 1994.