Acquisition of Translation Rules from Parallel Corpora

Yuji Matsumoto and Mihoko Kitamura

Graduate School of Information Science
Nara Institute of Science and Technology

Abstract

This article presents a method of automatic acquisition of translation rules from a bilingual corpus. Translation rules are extracted by the results of structural matching of parallel sentences. The structural matching process is controlled by the word similarity dictionary, which is also obtained from the parallel corpus. The system acquires translation equivalences of word-level as well as those of multiple word or phrase-level.

1 Introduction

The major issues in Machine Translation are the ways to acquire translation knowledge and to apply the knowledge to real systems without causing unexpected side-effect phenomena. Hand-coding of transfer rules suffers from the problems of enormous manual labour and the difficulty of maintaining their consistency.

Example-based translation (Sumita 90; Sato 90) is supposed to be a method to cope with this problem. Unlike transfer-based approaches, the idea is to carry out translation by referring to translation examples that give the best similarity to the given sentence. The key technique is to define the similarity between the given sentence and the examples and to identify the ones with the best similarity. Robustness and scalability are the claims of this approach. However, there are at least two important problems that haven't been answered. One is "knowledge access bottleneck," which concerns the selection of the most similar example. Similarities are usually defined only for fixed and local structures, such as predicate argument structures and compound nominals. The units of translation cannot always be such fixed structures and may vary according to the language pairs. Similarity should be defined in a more flexible way. The other is "knowledge acquisition bottleneck." In example based translation, the parallel examples have to be aligned not only at sentence-level but word or

phrase-level. Although the sentence-level alignment can be done automatically using statistics, e.g., (Utsuro *et al.* 94), the word-level alignment is not an easy task especially when the system tries to cover wide syntactic phenomena.

This paper presents a method of automatic acquisition of translation rules from a parallel corpus of English and Japanese. Translation rules in this paper refer to word selection rules and translation templates that represent word-level and phrase-level translation rules. A translation template are regarded as a phrasal translation rule. Since translation rules may change according to the target domain, this method shed a light on an easy and effective way for developing domain dependent translation rules by accumulating a parallel corpus.

2 Acquisition of Translation Rules

Figure 1 shows the flow of the acquisition of translation rules. Following three types of resources are assumed:

1. A Parallel corpus of the source and target languages.
2. Grammars and dictionaries of the source and target languages.
3. A machine readable bilingual dictionary.

The automatic acquisition of translation rules is composed of the following three processes:

Calculation of word similarities Calculation of the similarities of word pairs of the source and target languages based on their co-occurrence frequencies in the parallel corpus.

Structural matching Structural matching of the dependency structures obtained through parsing of parallel sentences.

Acquisition of translation rules Acquisition of translation rules based on the structural matched results.

We focus on a bilingual corpus of Japanese and English and assume that sentence-level alignment has been done on the corpus. In case they are not aligned, we can have them aligned using an existing alignment algorithm such as (Kay & Röscheisen 93) (Utsuro *et al.* 94).

2.1 *Calculation of Word Similarities*

We define the similarity of a pair of Japanese and English words by a numerical value between 0 and 1. We use the following two resources for

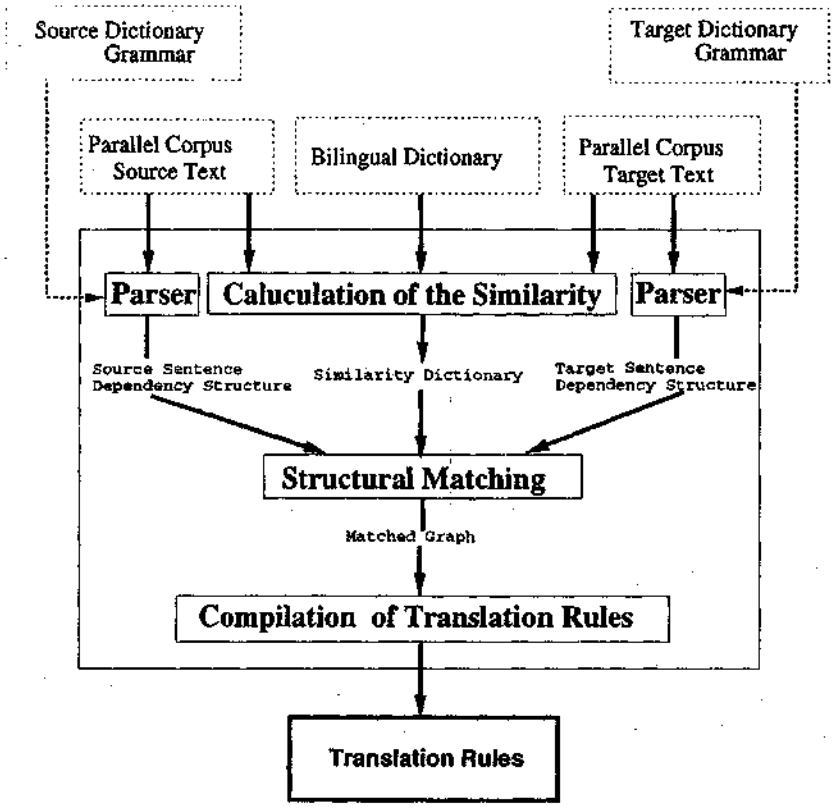

Figure 1: The flow of translation rules acquisition

obtaining the similarity:

- a machine readable bilingual dictionary
- a bilingual corpus of Japanese and English

As for the former, we assign value 1 to the translation pairs appearing in the bilingual dictionary. As for the latter, we use the basic calculation method of the similarity proposed by (Kay & Röscheisen 93). Unlike their method, we preprocessed the corpus by analyzing them morphologically to obtain the base form of the words. The similarity of a pair of Japanese and an English words is defined by the numbers of their total occurrences and co-occurrences in the corpus. The similarity of a Japanese and English

English: Companies compensate agents.
Japanese: 会社-は 代理店-に 報酬-を 与える。
The best score = 1.55

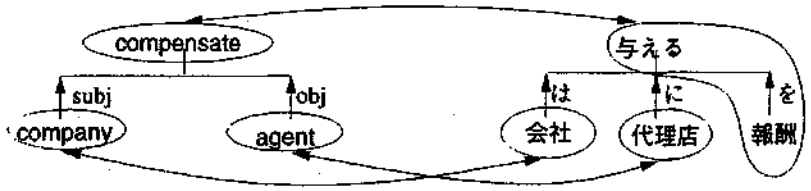

Figure 2: A result of structural matching

word-pair, is defined by $sim(w_J, w_E) = \frac{2f_{je}}{f_j + f_e}$, where f_j and f_e are the total numbers of the occurrences of the Japanese word w_J and the English word w_E, and f_{je} means the total number of co-occurrence of w_J and w_E, that is, the number of occurrences they appear in corresponding sentences.

2.2 *Structural matching of parallel sentences*

Corresponding Japanese and English sentences in the parallel corpus are parsed with LFG-like grammars, resulting in feature structures. We do not use any semantic information in the current implementation. When a sentence includes syntactic ambiguity, the result is represented as a disjunctive feature structure.

A feature structure is regarded as a directed acyclic graph (DAG). In the subsequent process of structural matching, we use the part of the DAG that relates with content words (such as nouns, verbs, adjectives and adverbs). The resulting DAG represents a (disjunctive) dependency structure of the content words in the sentence. We start with a pair of dependency graphs of Japanese and English sentences and find the most plausible graph matching between them. We use the word similarities described in the previous section in the matching process. The similarity of word pairs is extended to the similarity of subgraphs in the dependency structures. A sample result of structural match is shown in Figure 2.

The basic definition and algorithm follows (Matsumoto *et al.* 93), though the similarity measures of words and subgraphs are refined.

When the corresponding subgraphs (nodes in circles pointed by a bidi-

rectional arrow in Figure 2) consist of single words, the word similarity is used for their similarity. When any of the subgraphs contains more than one content word, we placed the following criterion: The higher the similarity of a word pair the finer their corresponding subgraphs should be. This means that mutually very similar words should have an exact match whereas mutually dissimilar words, when they are matched against each other by the structural constraint, are better included in coarse subgraphs. To achieve this criterion, we defined the following formula for calculating mutual similarity between subgraphs:

Let s and t be subgraphs matched against and V_s and V_t be the sets of contents words in s and t. We can assume, without loss of generality, that $|V_s|$ is not greater than $|V_t|$ (V_s and V_t can be switched if it is not the case).

Let D_p be the set of pairs of elements from $|V_s|$ and $|V_t|$ defined by an injection (one-to-one mapping) p: $|V_s| \rightarrow |V_t|$.

$$D_p = \left\{ \langle a, p(a) \rangle \mid a \in V_s \right\}$$

Then, the average similarity of words between $|V_s|$ and $|V_t|$ is defined as follows:

$$AverageSim = \frac{\max_p \left(\sum_{d \in D_p} sim(d) \right)}{|V_s|}$$

To achieve the above criterion, we put a threshold value Th ($0 < Th < 1$) where a similarity value higher than Th is supposed to indicate that they are mutually similar. The following formula of similarity between two subgraphs realizes the criterion in that the total similarity is bent toward the threshold value according to the size of subgraphs. Dividing the difference of $AverageSim$ and Th by the size of subgraphs works as a penalty for graphs that are mutually similar and as a reward for graphs that are mutually dissimilar.

$$sim(s,t) = \left(Th + \frac{AverageSim - Th}{|V_s| + |V_t| - 1} \right) \cdot |V_s|$$

The branch-and-bound algorithm is employed for the search of the graph matching that gives the highest similarity value. Figure 2 shows an example of dependency structures and the result of the structural matching, in which the corresponding pairs are linked by arrows. Here *the best score* is the total similarity of the most similar graph matching. The threshold is set at 0.15.

2.3 *Acquisition of translation rules*

After accumulating structurally matched translation examples, the acquisition of translation rules is performed in the following steps. We assume a thesaurus for describing the constraints on the applicability of the acquired rules. Suppose we concentrate on a particular word or a particular phrase in the source language graphs that appear as a subgraph in matching graphs. We refer to the subgraph as t.

1. Collect all the matched graphs that contain the same subgraph as t.
2. Extract the graph t and its children together with the corresponding part of the target language tree. Some heuristics are applied in this process: Corresponding pairs of pronouns are deleted, and zero personal pronouns in Japanese sentences are recovered.
3. The child elements are generalized using the classes in the thesaurus, which is identified as the condition on the applicability of the rule.

The system acquires two types of translation rules that represent word-level and phrase-level translation rules. When the top subgraph consists of a single content word, we regard that the corresponding subgraphs give a a word selection rule. On the other hand, when the top subgraph consists of more than one content word, we regard it as a phrasal expression, and call it a translation template. Figure 2 shows an example of phrasal-level correspondence, "compensate : 報酬-を 与える."

Since we assume the translation is influenced by the adjacent elements, i.e., the words that directly modify the word in the subgraph, we generalize the information in the collected matches so as to identify the exact contexts in which the translation rule is applicable.

From the set of partial graphs that share the same parent nodes, translation rules in the form of feature structures are obtained.

In the experiment described below, we focus on acquiring Japanese-English and English-Japanese translation rules related with verbs, nouns and adjectives.

3 Experiments of translation rule acquisition

We used *Torihiki Jouken Hyougenhou Jiten* (Collection of Japanese-English expressions for business contracts, 9,804 sentences) (Ishigami 92)

w_E	w_J	*Simirality*	f_e	f_j	f_{je}
abnormal	異常だ	1	2	2	2
accessory	付属品	0.923077	14	12	12
accountant	会計士	0.941176	9	8	8
accumulative	重畳	1	2	2	2
accurate	正確だ	0.769231	5	8	5
address	住所	0.764977	111	106	83
adjudge	要求額	1	2	2	2
administrative	残る	0.8	3	2	2
adopt	採択	1	2	2	2
advancement	進歩	1	4	4	4
advancement	品質管理	0.8	4	6	4
afterward	委ねる	0.8	2	3	2
agent	代理店	0.935583	1004	952	915

Table 1: Examples of word similarity

word	sentence	parsing	matching	word-level	phrase-level
与える	184	183(99.5%)	180(97.8%)	115(63.9%)	65(36.1%)
補償	254	245(96.5%)	242(95.3%)	144(59.5%)	97(40.1%)
有効だ	114	103(90.4%)	99(86.8%)	68(68.7%)	31(31.3%)
make	309	309(100%)	298(96.4%)	184(61.7%)	113(37.9%)
business	191	191(100%)	179(93.7%)	92(51.4%)	87(48.6%)
exclusive	127	127(100%)	116(91.3%)	27(23.3%)	88(75.9%)

Table 2: Statistics of parsing and matching results

and EDICT 1994[1] and *Kodansha* Japanese-English dictionary (Shimizu 79) (93,106 words) as the base resources. We also used an electronic version of Japanese thesaurus (called Bunrui-Goi-Hyo, BGH) (NLRI 94) and Roget's Thesaurus (Roget 11) for specifying the semantic classes. The current system works only with simple declarative sentences.

3.1 *Acquisition of translation rules*

Total of 948 word pairs of Japanese and English are obtained by the method for the calculation of word-word similarity between two languages described in Section 2.1. Some examples of the similarity obtained in the

[1]EDICT 1994 is obtainable through ftp via monu6.cc.monash.edu.au:pub/nihongo

experiment are shown in Table 1.

We get a number of domain specific terms about business contracts, such as "agent:代理店" and "accountant:会計士," which are not found in the ordinary bilingual dictionaries. Out of the 948 word pairs we obtained, only 236 appear in EDICT or *Kodansha* Japanese-English dictionary. Acquisition of word pairs from domain specific parallel corpora is very important, since many domain specific word pairs often do not appear in ordinary bilingual dictionaries. However, it should also be noted that the repetitive occurrences of the same expression causes a slight error in the similarity of the pairs.

We selected several Japanese and English words of frequent occurrence and collected structurally matched results. Some of the results for those words are shown in Table 2. For example, out of 184 occurrence of Japanese verb "与える", 183 sentences were successfully parsed (meaning that the correct parse was included in the possible parses), and 180 sentences succeeded in structural matching, in which 115 sentences had the top subgraph with a single content word, and 65 sentences had the top subgraph with more than one content word.

To acquire word selection rules, the results are classified into the groups according to the translated target words. A word selection rule is acquired from each target word by generalizing the child nouns by the classes in the thesaurus. The word selection rules for "与える" are summerized in the upper part of Table 3. For instance, the table specifies that "与える" is translated into "give" when its subject is either of the semantic classes, *substance*, *school*, *store* and *difference* and its object is either of the class of *difference*, *unit* and so on.

Phrasal translation rules are treated in the same way. Such examples of "与える" are shown in the lower part of Table 3. For instance, the Japanese phrase "X が Y に 報酬-を 与える" is translated into "X compensate Y", if X and Y satisfy the semantic constraints described in the table.

3.2 *The translation rules*

The translation rules described above are converted into the following data structure in our machine translation system.

```
tr_dict( index,
         source feature structure,
         target feature structure,
         condition).
```

Engllish verb	nominative(ga)	objective(wo)	dative(ni)
give(58)	[substance], [school],[store], [difference]	[difference],[unit], [chance],[feeling], [number],[start end]	[substance] [store],[school] [range seat track]
affect(8)	[change]	[cause]	[trade]
confer(6)	[school]	[propriety]	[store]
furnish(3)	[difference],[school], [store]	[school],[feeling]	[range seat track]
render(1)	[difference]	[care]	[range seat track]
afford(1)	[harmony]		
provide(1)	[difference]		

the number of word occurrence is in parentheses.
The name of semantic classes in the thesaurus is in square brackets.

Japanese patterns	English patterns
[1][store,school] が [2][store,school,cause,...] に 影響を 与える (17)	[1] affect [2]
[1][store,school] が [2][store,school] に報酬を 与える (2)	[1] compensate [2]
[1][store,school] が [2][store,school] に 同意を 与える (2)	[1] assent to [2]
[1][store] が [2][store] に 承認を 与える (1)	[1] authorize [2]
[1][store] が [2][store] に [3][substance] の必要量を与える (1)	[1] furnish [2] with [3]

the number of word occurrence is in parentheses.

Table 3: Acquired translation rules of "与える"

index The index word of the translation rule.
source feature structure A feature structure of the source language.
target feature structure A feature structure of the target language.
condition The semantic condition for the rule described by a set of semantic classes for the variables appearing in the source feature structure.

In the condition, `checksum/2` is a Prolog predicate for checking the semantic classes of the variables (semantics classes are expressed by the class numbers in the thesaurus). Identifying the most suitable semantic classes in the thesaurus is by no means an easy task. In the current implementation, we use the semantic classes at the lowest level in the Japanese thesaurus BGH, which has 6 layers.

This leads the description of the semantic condition to be a list of the lowest level semantic classes. Therefore, in our current implementation the translation rules compiled with few translation examples are far from

complete. Some of the final form of translation rules are represented as follows:

```
(1) 与える
tr_dict(与える, [ pred:与える(verb), が:X, を:[pred:同意(noun)],
                                        に:Z ],
                [ pred:assent(verb), subj:X,  to:Z ],
                true ).
tr_dict(与える, [ pred:与える(verb), が:X,  を:Y,  に:Z ],
                [ pred:give(verb), subj:X,obj1:Y, obj2:Z ],
                ( checksem(X,[11000,11040,11600,...]),
                  checksem(Y,[11642,11910,13004,...]),
                  checksem(Z,[11000,11040,12630,...]) ) ).
(2) 委託
tr_dict(委託, [ pred:委託(noun) ],
              [ pred:reference(noun) ],
               true ).
```

4 Discussion and Related Works

Our machine translation system based on the acquired translation rules has the following characteristics:

The system uniformly deals with word selection rules such as "confer:与える" and phrasal translation rules such as "X が Y に 報酬-を 与える: X compensate Y." Even if there is no translation rule to apply, the system uses the bilingual dictionary as the default. Translation pairs in the dictionary are regarded as word selection rules with no condition.

Since all the translation rules are acquired from translation examples, manual compilation of translation rules is made minimal. Also, since the structural matching results used to obtain the translation rules are symmetric, both English-Japanese and Japanese-English translation rules are acquired, making two-way translation possible.

Another important characteristic is that ambiguity (ambiguous translations caused by multiple applicable translation rules and ambiguous structural analyses) are resolved by putting priority to the translation rules with more specific information. The frequency information of translation pairs is also used for deciding the priority among the translation options.

The parsing and generation phases share the grammars and dictionaries that are used in the acquisition phase of the translation rules. This assures no contradiction among the parsing, generation and translation rules.

On the other hand, the following issues should be considered:

The quality of the translation rules depends on the quality of the thesaurus. There are some unadmissible word selection and phrasal rules acquired in the experiment. For example, the word selection rule, " X[human] に Y[problem] を 唱える ("唱える" means advocate)" was paired with "make Y[problem] to X[human]," which is not a good translation rule. Rather, "make an objection to X[human]: X[human] に異議を唱える" should be considered as an appropriate idiomatic expression. Idiomatic expressions like this example should be distinguished from normal word selection rules.

The proposed method is suitable to formal domains. An experiment with colloquial expressions reveals much more difficulties in acquiring "good" translation rules. Moreover, the current method cannot cope with expressions that necessitate contextual information.

The method should be augmented so as to deal with complex sentences. We do not think that a direct augmentation of the structure matching algorithm is applicable to complex sentences. Some two-level technique should be developed, the first level is to find an appropriate decomposition of complex sentences and the proposed structural matching is applicable at the second level.

A similar work for acquiring translation rules from parallel corpora is discussed in (Kaji 92), in which a bottom-up method is used for finding corresponding phrases (i.e. partial parse trees). We use dependency structures, which we think, is a critical point, since word order is not normally preserved between Japanese and English sentences while dependency between content words is preserved in most of the cases.

(Watanabe 93) proposed a method of using matched pairs of dependency structures of Japanese and English sentences for improving translation rules. The algorithm of finding the structural correspondence is different from ours. Our method uses a more finer similarity measure that is learned from parallel corpus. As for the translation rule acquisition, their objective is to improve existing transfer rules whereas our objective is to compile the whole translation rules altogether.

5 Conclusions

The translation rules obtained by the proposed method can be integrated into an existing machine translation system. Generally, translation may differ depending on the domain. Our system is easily adapted to any domain provided that sizable parallel corpora of that domain are accumulated.

To improve the acquired translation rules both in quality and quantity, we need to enlarge the scale of the parallel corpora. Another possible way to improve the translation rules is to give the post-edited translation results back to the acquisition phase. By doing this, missing translation rules are gradually acquired.

REFERENCES

Ishigami, Susumu. 1992. *Torihiki Jouken Hyougenhou Jiten.* Tokyo: International Enterprise Development Co.

Kaji, Hiroyuki, Y. Kida & Y. Morimoto. 1992. "Learing Translation Templates from Bilingual Text". *Proceedings of the 14th International Conference on Computational Linguistics (COLING-92)*, vol.II, 672-678. Nantes, France.

Kay, Martin & M. Röscheisen. 1993. "Text-Translation Alignment". *Computational Linguistics* 19:1.121-142.

Matsumoto, Yuji, H. Ishimoto & T. Utsuro. 1993. "Structural Matching of Parallel Texts". *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL'93)*, 23-30. Columbus, Ohio.

National Language Research Institute. 1994. *Bunrui-Goi-Hyo [Word List by Semantic Principles].* Tokyo: Syuei Syuppan.

Roget, Peter M. 1911. *Roget's Thesaurus.* New York: Crowell.

Sato, Satoshi & M. Nagao. 1990. "Toward Memory-Based Translation". *Proceedings of the 14th International Conference on Computational Linguistics (COLING-90)*, vol.III, 247-252. Helsinki, Finland.

Shieber, Stuart M., G. van Noord, R.C. Moore & F.C.N. Pereira. 1990. "A Semantic Head-Driven Generation Algorithm for Unification-Based Formalisms". *Computational Linguistics* 16:1.30-42.

Shimizu, Mamoru & N. Narita. 1979. *Japanese-English Dictionary.* Tokyo: Kodansha Co.

Sumita, Eiichiro & H. Iida. 1991. "Experiments and Prospects of Example-Based Machine Translation". *Proceedings of 29th Annual Meeting of the Association for Computational Linguistic (ACL'91)*, 185-192. Berkeley, California.

Utsuro, Takehito, H. Ikeda, M. Yamane, Y. Matsumoto & M. Nagao. 1994. "Bilingual Text Matching Using Bilingual Dictionary and Statistics". *Proceedings of the 14th International Conference on Computational Linguistics (COLING-94)*, vol.II, 1076-1082. Kyoto, Japan.

Watanabe, Hideo. 1993. "A Method for Extracting Translation Patterns from Translation Examples". *Proceedings of the 5th Int. Conf. on Theoretical and Methodological Issues in Machine Translation (TMI-93)*, 292-301. Kyoto, Japan.