ACQUISITION OF TRANSLATION RULES FROM PARALLEL CORPUS

Yuji Matsumoto **Mihoko Kitamura**
Graduate School of Information Science
Nara Institute of Science and Technology
Takayama, Ikoma, Nara 630-01, Japan
{matsu, mihoko-k}@is.aist-nara.ac.jp

Abstract

Machine translation is one of the important applications of natural language processing techniques and has been worked out for practical use. However, there is the problem of how we should acquire the information needed for translation, such as grammar rules and lexical rules.

This paper presents a method for automatic acquisition of lexical information for machine translation, especially lexically dependent translation rules. The acquisition process uses the similarity measure of word pairs obtained from a machine readable bilingual dictionary and the one obtained statistically from parallel corpus.

Pairs of translation examples are syntactically analyzed and are represented as (disjunctive) dependency graphs. Structural matching is done based on the similarity measure between subgraphs that is defined from the similarity between word pairs. Two types of translation rules are then acquired according to the types of the matching results. The rule types are: 1) word selection rules, and 2) fused or idiomatic translation rules.

The experimental result using 36,000 sentences shows that this method is useful for obtaining rules and patterns for translation.

1 Introduction

Machine Translation has a long history and has been most practically studied in the field of natural language processing. The major problems are how to build or acquire the knowledge of translation, or translation rules, and how to apply such knowledge to real systems without suffering from unexpected side-effect phenomena.

One proposal to cope with those problems is, unlike traditional rule based approaches to translation, to use translation examples directly without relying on explicit translation rules, called example based translation [Sato 90] [Sumita 90]. The idea behind this method is to make translation by referring to the sample translation that gives the best similarity to the given sentence. The key technique is to define the similarity and to identify examples with the best similarity.

The current example based translation systems assume the examples be in a certain syntactic structure, such as argument structures of verb phrases or PP-modified noun phrases. They also assume that the examples are syntactically analyzed beforehand. The important problems will be that it is not easy to collect such syntactically analyzed examples and that it is not flexible to confine input sentences be in a certain syntactic patterns. Example based translation should be adaptive to far broader types of syntactic structures. Furthermore, a systematic way to accumulate analyzed examples should be investigated.

This paper presents a method to acquire translation rules in a more flexible sense from a parallel corpus. Translation rules in this paper refer to word selection rules and translation templates that represent phrasal or multi-phrasal translation rules. A translation template can be regarded as an idiomatic translation rules. However, as translation rules, the sense of idiomatic expression varies from language to language. Moreover, even in a single language, idiomatic expressions have to have different definition in some specific domain.

There have been a number of works for sentence alignment of bilingual texts [Brown, et al. 91] [Gale et al. 91] [Kay 93], giving a very good results in finding correspondence at sentence level. We assume in this paper that sentence level alignment has been done in bilingual corpora. We currently have an aligned Japanese-English translated sentences extracted from a Japanese-English dictionary.

We proposed a method of acquiring translation rules based on a technique of structural matching of parallel sentences. Parallel corpus is first used to acquire similarity measure of word pairs between two languages. Then, parallel sentences are structurally matched according to the similarity measure. Using a thesaurus as a semantic classification, translation rules are constructed from the matched results.

2 Matching Parallel Sentences

We assume that sentence-level alignment has been done on the parallel corpus. We first define the similarity measure between word pairs of the two languages, and then show the structural matching of corresponding sentences, from which translation rules are acquired. We focus on a bilingual corpus of Japanese and English.

2.1 Word-word similarity between two languages

There are two sources for obtaining similarity measure between word pairs of Japanese and English: 1) machine readable bilingual dictionaries, and 2) the bilingual corpus itself. We define the similarity of a pair of Japanese and English words as a numeral value between 0 and 1. At present, we put value 1.0 to the translation pairs obtained from machine readable dictionaries:

$$sim(\langle w_J, w_E \rangle) = 1 \qquad (\langle w_J, w_E \rangle \in Tp)$$

where Tp is the set of translation pairs taken from bilingual dictionaries.

Many word pairs that are in some sense similar often do not appear in bilingual dictionaries simply because they are not exact translation. Some flexible measure of similarity between word pairs are indispensable in finding out correct structural matching that is shown in the next section. Furthermore, word-word similarity varies according to the domain of the corpus. We, therefore, extract word-word similarity from the bilingual corpus itself. The basic calculation method of the similarity is taken from [Kay 93]. Unlike their method, we preprocessed the corpus by analyzing it morphologically to obtain the base form of the words. We did not take their part-of-speech in consideration since there is not necessarily a clear correspondence of part-of-speech between Japanese and English words.

The similarity of a pair of a Japanese word and an English word is defined by their total occurrences and co-occurrences in the corpus as follows:

$$sim(\langle w_J, w_E \rangle) = \frac{2 f_{je}}{f_j + f_e}$$

where f_j and f_e are the total numbers of occurrences of the Japanese word w_J and the English word w_E, and f_{je} means the total number of co-occurrences of w_J and w_E, that is, the number they appear in corresponding sentences. Some example of the similarity measure taken from our bilingual corpus is shown in Table 1.

2.2 Disjunctive graph representation of sentences

A corresponding pair of Japanese and English sentences are parsed with LFG-like grammars resulting in feature structures. We do not use any semantic information, and solely syntactic

w_J	w_E	$sim(w_J, w_E)$	f_j	f_e	f_{je}
わかる	know	0.1659292	359	545	75
わかる	understand	0.22921348	359	86	51
ダイヤ	schedule	0.15384615	10	16	2
ダイヤ	diamond	0.69565217	10	13	8
チーズ	cheese	0.8	2	3	2
チーム	team	0.90434783	55	60	52
チップ	tip	0.22222222	2	16	2
テープ	tape	1	3	3	3

Table 1: A sample list of similarity measure

information is used for the analysis. When a sentence involves syntactic ambiguity, the result is represented as a disjunctive feature structure.

A feature structure can be regarded as a directed acyclic graph (DAG). In the subsequent process of structural matching, we use the part of the DAG that relates content words (such as nouns, verbs, adjectives and adverbs). The resulting DAG reflects the dependency relation between the content words in the sentence. We start with a pair of (disjunctive) dependency graphs of Japanese and English sentences and find out the most plausible graph node matching between them. The plausibility is defined from the similarity measure of words appearing in the sentences.

2.3 Similarity between subgraphs

The similarity between word pairs is extended for defining the similarity between subgraphs in the graph representation of parallel sentences. Although the basic definition and algorithm is taken from [Matsumoto et al. 93], the similarity measure of words and subgraphs are substantially redefined.

The similarity between two subgraphs is defined by the sets of content words appearing in them. Let s and t be subgraphs of the dependency graphs of Japanese and English sentences, and V_s and V_t be the sets of content words in s and t. Assume without loss of generality that the size of V_s is not larger than that of V_t. We interchange V_s and V_t when $\mid V_s \mid > \mid V_t \mid$.

Taking an arbitrary injection p from V_s into V_t ($p : V_s \rightarrow V_t$), a set of pairs is defined by each of such p's.

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

Then, we define the average similarity of words in the subgraphs as the average of the maximal value of possible matching between words in the subgraphs:

$$AverageSim(\langle s,t\rangle) = \frac{\max_p\left\{\sum_{d\in D_p} sim(d)\right\}}{\mid V_s \mid}$$

In the overall matching of the dependency graphs of corresponding Japanese and English sentences, words with high similarity should get a fine grained matching while words with low similarity should not get matched. In case a pair of words with low similarity are to be matched due to syntactic constraint, they should better get matched coarsely. The reason is that while a pair of words with low similarity are rarely matched at the word level they may be matched at the phrasal level, especially in idiomatic expressions.

To define the similarity between subgraphs, we pre-determine a threshold value of similarity. Then, a pair of words with a similarity greater than the threshold are considered really similar, and the one with a similarity less than the threshold are considered dissimilar. By referring to the threshold value as Th[1], the similarity between the subgraphs s and t is defined by the following expression:

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

The branch-and-bound algorithm is employed for searching the subgraph matching that gives the highest similarity value as the whole, the detail of which is omitted for the space limitation. Some experimental results are found in Figure 1. Dotted circles show subgraphs in the dependency graphs and bidirectional arrows show the matches between subgraphs. In the upper example, a precise word level matching is obtained, while a subgraph matching is found in the lower.

3 Acquisition of Translation Rules

3.1 Types of translation rules and method of acquisition

After accumulating structurally matched translation examples, translation rules are acquired by generalizing the examples. We are working on acquiring two types of translation rules. In either case, we focus on a subgraph together with its peripheral subgraphs and their corresponding subgraphs.

When the top subgraph consists of a single content word, we regard that the corresponding subgraph shows a possible translation of the word. For instance, the upper graph of Figure 1 shows that a possible translation of the word “かける” is “put” in this case.

---

[1] In the experiments we set $Th = 0.1$.

English: I put a kettle on the stove.
Japanese: 私 - は ストーブ - に やかん - を かける
I - TOP stove - LOC kettle - ACC 'kakeru'

English: I call him.
Japanese: 私 - は 彼 - に 電話 - を かける
I - TOP him - DAT telephone - ACC 'kakeru'

Figure 1: Samples of structural matching

On the other hand, when the top subgraph contains more than one content word, we regard it as a fused, or idiomatic, expression. The lower example in Figure 1 shows that the phrase “電話 を かける” can be translated into a single English word “call”.

Of course, the translation is influenced by the peripherals, i.e., the words that directly connected with the subgraph. We try to generalize the peripheral information from the collected matches so as to identify the exact contexts in which a certain translation rule is effective.

Accordingly, we make a simple categorization of the translation rules: 1) the top subgraph consists of a single word, and 2) the top subgraph consists of more than one word. The former cases are used to acquire word selection rules, and the latter are used to acquire fused or idiomatic translation rules.

In the next section we focus on acquiring Japanese-English translation rules related with Japanese verbs and verb phrases. Since the structural matching is irrelevant to the direction of translation, the same technique is usable to acquire English-Japanese translation rules. Also, note that any part in the matched examples are usable for rule acquisition. Acquisition of word selection rules for nouns or noun phrases can be done by concentrating on the subgraphs headed by nouns.

3.2 Experiment of translation rule acquisition

We make use of an aligned corpus of about 36,000 parallel sentences obtained from a Japanese-English dictionary. Most of the sentences are simple declarative sentences.

For the bilingual dictionary to get the similarity of word pairs we used EDICT (V94-002) compiled by James Breen[2]. The similarity value 1.0 is given to the pairs appearing in the dictionary. We also used the corpus to obtain similarity values of word pairs that do not appear in EDICT.

We focused on several frequently occurring Japanese verbs and accumulated the structurally matched results. For example, Japanese word “出る (show up, move outside, come out from some source)” appeared 241 times in the corpus, of which 79 are matched against a single English verb. The results are classified into the groups according to the English verbs, from each of which a word selection rule is acquired by generalizing the peripheral nouns together with the case markers. A part of the results is shown in Table 2. We made use of an electronic version of a Japanese thesaurus, called Bunrui-Goi-Hyo [NLRI 64], for classifying Japanese nouns appearing at case positions of the verb. The numbers in the table represents a class in the thesaurus. An English explanation is put in parentheses for the sake of readability. For instance, the table shows that when Japanese verb “出る” is accompanied

---

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

English verb (occurrence)	nominative (ga)	objective (wo)	dative (ni)
leave (22)	12000(human) 14650(vehicle)	12540(city, colony) 144xx(residence) 14700(land, plot)	
go (out) (28)	12000(human) 14650(vehicle)	12540(city, colony) 12660(house, room) 144xx(residence)	12600(society) 13371(trip) 13800(business) 15201(transit)
appear (9)	12000(human) 15150(climate) 15251(celestial body)		1525x(sky)
rise (5)	151xx(climate) 15251(celestial body)		

Table 2: Word selection rules for “出る”

with a human or a vehicle as its subject and a city, etc as its object, it can be translated into the English verb “leave”.

Though other translation rules in which more than one English word are involved can be dealt with in the same way, it was very rare that same translation patterns occur in such cases. A few such examples are: Japanese phrase “裏目に出る (produce a result opposite to one’s plan)” is translated into “backfire”, and “幽霊が出る (ghost appears)” is translated into “be haunted”.

The size of current corpus is far unsatisfiable to acquire complex translation rules. However, some very interesting rules have been acquired. An example is shown in Figure 2, in which a large fused body is obtained as a translation rules. Note that the rule represents the same meaning whereas the words in Japanese and English rules are semantically irrelevant at all. (The literal meaning of the Japanese sentence is ‘there is not a person who goes beyond to the right of him in eloquence.’)

3.3 Discussion on the experiments

This section gives conclusions and discussion with respect to the experiments. As for word selection rules, we have learned that when the same translation patterns appear in the corpus more than a certain number, say more than 20 times, quite plausible rules are acquired. Although much larger corpus is necessary to acquire complex translation rule, some interesting rules have been acquired.

English:	He is second to none in eloquence.					
Japanese:	弁舌-では	彼-の	右-に	出る	者-は	いない
	eloquence	his	right	go beyond	person	not-exist

Figure 2: Sample of complex translation rule

There are a number of problems to be investigated. Semantics classification in exisisting thesauri is not powerful enough to describe precise translation rules. This is the general problem in semantics treatment of corpus analyses. Besides, even in word selection rules, idiomatic expressions should be distinguished from normal usages. At present, we do not make any distinction between idiomatic and non-idiomatic expressions. We simply take the class of nouns at the case positions of verbs as small as possible in the thesaurus.

In the case of complex translation rules, some portion of a rule has to be taken as a variable or to be dependent on some other part in the sentence. In the example of Figure 2, though the English expression "be second to none" is a fused idiomatic expression, the word "者 (person)" in the Japanese expression should vary according to the semantic category of the noun at the position of "彼 (he)". For instance, "France is second to none in cuisine" should be translated into "料理ではフランスの右に出る 国 (country) はない".

The corpus used in the experiments contains sentences taken from a Japanese-English dictionary, in which most of the sentences are simple sentences. We should extend our technique to be applicable to more complex sentences. Or, we need to pre-process the corpus at the phrase level. We hope that our similarity measure of subgraphs is usable for finding phrase-level alignment of parallel corpora.

4 Conclusions

Before giving conclusions, we should comment on some related works. Similar attempts have been done by several authors. Kaji, et al, [Kaji et al. 92] attempts to learn translation templates from examples. Their work does not describe the word similarity and phrase similarity in a precise way. The key difference compared with ours is that they assume phrase-phrase correspondence in translation sentences. As can be found in Figure 2, the correspondence should not be restricted to phrases. Our method provides more freeness in the structural correspondence. An acquisition of translation rules is proposed in [Watanabe 93]. In his method, extraction of translation patterns is triggered by a translation mismatch between a translation example and the translation result of a MT system. Though the idea is very interesting, the approach is quite different from ours.

In this paper, we have shown a method of acquiring translation rules from a parallel corpus based on a precise structural matching of translation examples. Two types of translation rules are acquired. One is word selection rules, and the other is fused or idiomatic translation rules. Although lots of works should be done especially for the latter case, useful results have been obtained using a small corpus with about 36,000 translation examples.

References

[Brown, et al. 91] Brown, P.F., Lai, J.C. and Mercer, R.L., 1991. Aligning Sentences in Parallel Corpora, *ACL-91*, pages 169–176.

[Gale et al. 91] Gale, W.A. and Church, K.W., 1991. A Program for Aligning Sentences in Bilingual Corpora, *ACL-91*, pages 177–184.

[Kaji et al. 92] Kaji, H., Kida, Y., and Morimoto, Y., 1992. Learning Translation Templates from Bilingual Text, *COLING-92*, pages 672–678.

[Kay 93] Kay, M., and Röscheisen, M., 1993. Text-Translation Alignment, *Computational Linguistics*, Vol.19, No.1, pages 121–142.

[Matsumoto et al. 93] Matsumoto, Y., Ishimoto, H. and Utsuro, T., 1993. Structural Matching of Parallel Texts, *ACL-93*, pages 23–30.

[NLRI 64] NLRI, National Language Research Institute 1964. Word List by Semantic Principles, 'Bunrui-Goi-Hyo' (in Japanese), *Syuei Syuppan.*

[Sato 90] Sato, S. and Nagao, M., 1990. Toward Memory-Based Translation, *COLING-90*, Vol.3, pages 247–252.

[Sumita 90] Sumita, E., Iida, H. and Kohiyama, H., 1993. Translating with Examples: A New Approach to Machine Translation, *TMI-90.*

[Utsuro et al. 93] Utsuro, T., Matsumoto, Y. and Nagao, M., 1993. Verbal Case Frame Acquisition from Bilingual Corpora, *IJCAI-93*, pages 1150–1156.

[Watanabe 93] Watanabe, H., 1993. A Method for Extracting Translation Patterns from Translation Examples, *TMI-93*, pages 292–301.