Learning English Verb Selection Rules from Hand-made Rules and Translation Examples

Yasuhiro AKIBA† Megumi ISHII†
Hussein ALMUALLIM‡ Shigeo KANEDA†

†) NTT Communication Science Laboratories.
1-2356, Take, Yokosuka-shi, Kanagawa-ken,
238-03, Japan.
E-mail: {akiba, megumi, kaneda}@nttkb.ntt.jp

‡) King Fahd Univ. of Petroleum & Minerals.
P.O. Box 801, Dhahran 31261, Saudi Arabia.
E-mail: hussein@ccse.kfupm.edu.sa

Abstract

This paper proposes a method based on machine learning that automatically acquires English verb selection rules for machine translation. If the rules are learned from only real translation examples, many examples are necessary for good translation quality. It is, however, difficult to gather a sufficiently large number of real translation examples. The main causes are verbs of low frequency and the frequent usage of the same sentences. To resolve this problem, the proposed method learns English verb selection rules from hand-made translation rules and a small number of real translation examples. The proposed method has two steps: generating artificial examples from the hand-made rules, and then putting as a training set, the resultant artificial examples and real examples into an internal learner. To evaluate the validity of the proposed method, English verb selection rules of NTT's Japanese-English Machine Translation System ALT-J/E are experimentally learned from hand-made rules and real examples. The resultant rules have better accuracy than either those constructed from the real examples or those that are hand-made.

1 Introduction

This work aims at the automatic acquisition of semantic analysis rules for rule-based Japanese-English Machine Translation (MT) systems. To realize the aim, this paper proposes an automatic acquisition method of English verb selection rules.

The rule-based Japanese-English MT system called *"ALT-J/E"* is being developed at Nippon Telegraph and Telephone Corporation (NTT) [Ikehara et al. 1989, Ikehara et al. 1990]. ALT-J/E has about fifteen thousand hand-made English verb selection rules, semantic analysis rules, in its Japanese-English translation knowledge base. To improve the translation quality of ALT-J/E, more English verb selection rules and more specific English verb selection rules are needed for each translation domain. Unfortunately, this would require excessive effort using the conventional hand-made approach. Thus, rules should be acquired automatically.

Almuallim introduced two algorithms to learn English verb selection rules from Japanese-English translation examples [Almuallim 94c]. These algorithms need many translation examples to learn good rules. It is very difficult, however, to gather many real translation examples from existing documents because some sentences are used repeatedly while a large number of verbs occur infrequently.

Thus, to overcome this scarcity of real examples in existing documents, some kind of information should be extracted from human knowledge. One practical way to extract human knowledge is to write hand-made English verb selection rules. These hand-made rules are, however, not complete nor sufficient for practical use. If these rules and real examples could be integrated, better translation performance would be obtained automatically. However, no such algorithm has been published up to now.

This paper proposes a new method that learns English verb selection rules with high accuracy from hand-made rules and sparse real examples. The proposed method generates examples from hand-made English verb selection rules. The examples are called *"artificial examples"* hereafter. The artificial examples and "real examples" are put into an internal learner. The internal learner can be any attribute-based learner.

To represent the importance of artificial examples and real examples, a weighting is given to each example in the proposed method. If a hand-made rule is very accurate, the artificial examples generated from it should be assigned a large weighting. On the other hand, if the hand-made rule is inaccurate, only a small weight is assigned. The proposed method determines the optimum weighting by cross validation.

To estimate the validity of the proposed method, English verb selection rules of the ALT-J/E system are experimentally learned from hand-made rules and real examples gathered from an existing document [Horiguchi 89]. In this experiment, the internal learner is Almuallim's learning algorithm using ID3 (C4.5) [Almuallim 94c]. The English verb selection rules learned by the proposed method have better accuracy than either those constructed from only real examples or those that are hand-made.

A brief explanation of English verb selection rules is given in section 2. Section 3 surveys a conventional algorithm and describes its problem. The new learning method is proposed in section 4. Experimental results are shown in section 5, and section 6 concludes this paper.

2 English Verb Selection Rules

This section details English verb selection rules and discusses automatic rule acquisition.

2.1 English Verb Selection Rules

This paper defines an English verb selection rule as having a Japanese pattern as its left-hand side and an English verb as its right-hand side, as shown in Figure 1. Such rules associate a Japanese pattern with an English verb. Here, a Japanese pattern consists of only one Japanese verb and the variables N_1, N_2, etc., which represent various Japanese sentence components, such as the Subject and the Object; "Fish", "Seafood", etc. are semantic categories. ALT-J/E has about 3,000 semantic categories that constitute a semantic hierarchy with 12 levels. Figure 2 shows a part of the semantic hierarchy.

ALT-J/E has a semantic dictionary with 400,000 words that are nouns or proper nouns. The semantic dictionary maps each Japanese noun to its appropriate semantic categories. Note that a noun usually has plural semantic categories. For example, the semantic dictionary states that the noun 鶏

			THEN	
IF			THEN	
	J-Verb	= “焼く (yaku)”		
	N_1 (Subj)	≡ “People”	E-Verb	= “bake”
	N_2 (Obj)	≡ “Bread” or “Cake”		
IF			THEN	
	J-Verb	= “焼く (yaku)”		
	N_1 (Subj)	≡ “People”	E-Verb	= “roast”
	N_2 (Obj)	≡ “Meat”		
IF			THEN	
	J-Verb	= “焼く (yaku)”		
	N_1 (Subj)	≡ “People”	E-Verb	= “broil”
	N_2 (Obj)	≡ “Fish” or “Seafood”		
IF			THEN	
	J-Verb	= “焼く (yaku)”		
	N_1 (Subj)	≡ “Agents”	E-Verb	= “cremate”
	N_2 (Obj)	≡ “People” or “Animals”		
IF			THEN	
	J-Verb	= “焼く (yaku)”		
	N_1 (Subj)	≡ “Agents” or “Machines”	E-Verb	= “burn”
	N_2 (Obj)	≡ “Places” or “Objects” or “Locations”		

where “ ≡ ” indicates “an instance of”.

Figure 1: English verb selection rules for the Japanese verb 焼く (yaku).

(niwatori), which means “chicken” or “hen” in English, is an instance of the categories “Meat” and “Birds”.

2.2 Relation between English Verb Selection Rules and Semantic Analysis

In order to show how English verb selection rules are semantic analysis rules, matching a Japanese sentence to English verb selection rules is described below. For example, when the Japanese sentence “コックがアップルパイを焼く[1](Kokku ga appurupai wo yaku.)” is input, ALT-J/E runs morphological analysis and syntactic analysis and analyzes the sentence into the Japanese verb “焼く (yaku)”, the noun “コック (kokku)”, and the noun “アップルパイ (appurupai)”. ALT-J/E states that “コック (kokku)” has 3 semantic categories, “tools”, “people”, and “jobs”, while “アップルパイ (appurupai)” has only one semantic category “confectionery” .

[1]This Japanese sentence means “A cook bakes an apple-pie.”.

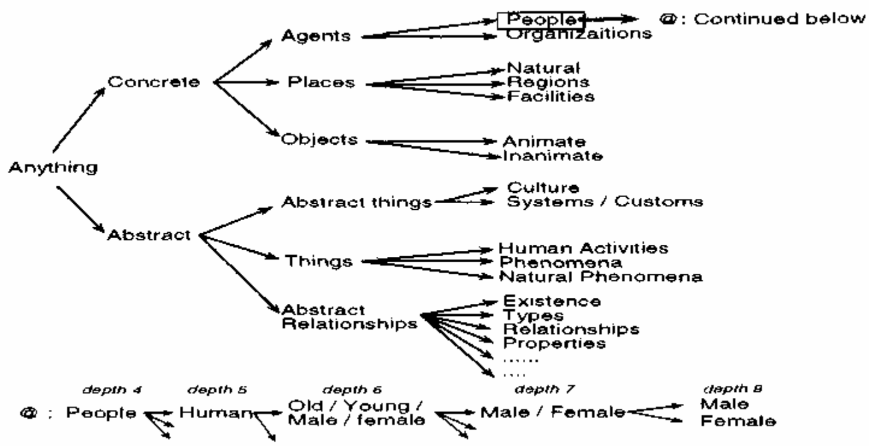

Figure 2: The upper levels of the Semantic Hierarchy in ALT-J/E.

Matching succeeds when each noun in the sentence has a descendant, on the semantic hierarchy, of one case element in the rule. In this example, the Objective case in the sentence (appurupai) matches only the first rule in Figure 1, while the Subjective case in the sentence (kokku) matches only those rules that contain semantic category "people". Therefore, only the first rule for the English verb "bake" satisfies both Subjective case (N_1) and Objective case (N_2). Thus the noun "コック (kokku)" means "people" not "tools".

As shown above, matching with English verb selection rules works by semantically analyzing Japanese sentences for MT.

2.3 Difficulty of acquiring English Verb Selection Rules

The most difficult task in acquiring English verb selection rules is to select semantic categories to each case element in the rules because this involves a huge number of combinations of the nearly 3000 semantic categories of ALT-J/E. Therefore, English verb selection rules should be acquired automatically. The automatic acquisition of English verb selection rules is the subject of learning algorithms.

3 The Former Work and its Problem

This section introduces the assumption of a former approach and confirms its infeasibility. Almuallim proposed two algorithms to learn English verb selection rules from Japanese-English translation examples [Almuallim 94c]. In his approach, training examples are prepared through the following process:

(1) Prepare pairs of a simple sentence and an appropriate English verb like the following:

("コックがアップルパイを焼く" "bake"),

(2) Parse the Japanese sentence in each pair,

(3) Pick up head nouns, and

(4) Make training examples like the following from the nouns as in step (3):

⟨ N1(Subj) ≡ "tools", "people", or "jobs",
N2(Obj) ≡ "confectionery",
"bake" ⟩,

where " ≡ " indicates "an instance of".

His approach needs many training examples to construct English verb selection rules that offer high accuracy. To investigate whether or not it is possible to get enough training examples to construct English verb selection rules with high accuracy, a corpus with about 50,000 Japanese-English translation entries was formed from existing documents [Keene 91, Cultural Affairs 90]. The 50,000 Japanese simple sentences contain about 5,000 different Japanese verbs. Only 1% of all Japanese verbs were used in 100 or more Japanese sentences. The translation examples contain repeated sentences.

Because approximately 100 sentences are needed per verb to ensure sufficient accuracy and 95% of all the Japanese verbs were used in 2 or more Japanese sentences, the analysis shows that, at least, about 2.5 million translation examples are required to construct good English verb selection rules for 95% of the 5,000 verbs. This number of translation examples is too huge to permit them to be gathered.

In our opinion, it is too optimistic to think that any learning algorithm can construct English verb selection rules from corpora extracted from just existing documents. That is, we need a new algorithm to construct English verb selection rules from hand-made English verb selection rules and real translation examples. The algorithm must offer adequate performance even if the number of examples is not enough to construct good rules and the hand-made rules don't have enough quality for practical use. In the next section, a learning method will be proposed that realizes the above approach.

4 Revision Learner

This section proposes the new method called *"Revision Learner"* that composes English verb selection rules from hand-made rules and real translation examples. Before talking about Revision Learner, we will clarify the learning task.

4.1 Learning Task

For a given Japanese verb *J-verb* and a possible English translation $E\text{-}verb_i$ of that verb, the algorithm has to find the appropriate condition(s) that should hold in the context in order to map *J-verb* to $E\text{-}verb_i$.

To learn English verb selection rules for a Japanese verb, for example the Japanese verb 焼く (yaku), the Learning task is described as the following:

【Learning Task】

Step-I Make English verb selection rules by hand[2], as shown in Figure 1,

Step-II Gather real examples[3], as (4) in section 3, and

Step-III Automatically construct the final rule from the above examples and rules by the Revision Learner mentioned in the next subsection.

[2]The rules don't have enough accuracy.

[3]The number of which is not enough to construct the English verb selection rules.

4.2 Outline of Revision Learner

Revision Learner can get information from hand-made English verb selection rules and real examples. If a hand-made rule is very accurate, Revision Learner needs to strongly weight the hand-made rule. On the other hand, if the hand-made rule is inaccurate, only small weighting should be assigned to the rules.

For weighting hand-made rules and real examples, Revision Learner uses numerical values called *"weighting values"* that is prepared in advance. In general, Revision Learner cannot know the optimum values. Revision Learner determines the optimum values from the given candidate weighting values. When the number of candidate weighting values is N, Revision Learner is outlined below:

【Revision Learner】

Step-i Generate examples from the hand-made English verb selection rules, where the details are described in subsection 4.3,

Step-ii Form a family of example sets $\{Data_i; i = 1 \cdots N\}$, where $Data_i$ is the union set of the artificial examples and the real examples with the ith candidate weighting values,

Step-iii For each $Data_i (i = 1 \cdots N)$, calculate average accuracy A_i of a rule learned from $Data_i$, by using cross validation[4],which is described in subsection 4.4, and

Step-iv Finally, output the rule that has the best average accuracy in $A_i (i = 1 \cdots N)$.

4.3 Artificial Example Generation Method

This section details Step-i in section 4.2, Artificial Example Generation. The Artificial Example Generation Method is outlined below:

Step-A Decompose the hand-made English verb selection rules into unit rules, here a unit rule is one like the following rule[5]:

[4]When cross-validation is executed, the weighting value of test examples should be 1.0.

[5]Negation in the condition part is expressed like the following form: N_1 = not V_1.

$$\text{IF } (N_1 \equiv V_1) \ \& \ (N_2 \equiv V_2) \ \& \cdots \text{ THEN Class} = \text{CV},$$

where " $\equiv$ " indicates "an instance of", N_1, N_2 etc. are case elements, V_1, V_2 etc. are semantic categories, and CV is an English verb,

Step-B From the above unit rules, generate artificial examples in the following form:

$$\langle N_1 \equiv v_1, N_2 \equiv v_2, \cdots, CV \rangle$$

where " $\equiv$ " indicates "an instance of", N_1, N_2 etc. are case elements, v_i is randomly selected from the descendant of V_i on the semantic hierarchy and V_i is a semantic category in the unit rule, and

Step-C Repeat Step-B until the desired number of examples are generated.

4.4 Cross Validation

This section details Step-iii in section 4.2, cross validation, strictly speaking m fold cross validation. Cross validation is well known in the Machine Learning. While it is usually used just to estimate the accuracy of a rule in experiment, our approach employes it in the learning stage.

To simplify the explanation, each $Data_i$ in section 4.2 Step-iii is expressed by the data set D. Given integer m and data set D, cross validation is outlined below:

Step-a Make m subsets S_k, which are disjoint, of the given set D,

Step-b From each difference set $D \setminus S_k (k = 1 \cdots m)$, learn a rule $rule_k$ by using the internal learner,

Step-c For each $rule_k (k = 1 \cdots m)$, calculate the accuracy $accuracy_k$ by using the remainder of set S_k as test examples, and

Step-d Calculate the average of $accuracy_k (k = 1 \cdots m)$.

Note that it is well known among Machine Learning researchers that integer m should be 10 or the number of elements in set D to estimate the accuracy of a rule.

5 Experimental results

The proposed method was tested in experiments. The applied internal learner was Almuallim's learning method [Almuallim 94c] which uses machine learning algorithm ID3 (C4.5) [Quinlan 86, Quinlan 92].

5.1 The Experiment

The experiment is described below:

Evaluation Data In this evaluation, hand-made rules were selected from the English verb selection rules in ALT-J/E. Real examples were made with reference to an existing document [Horiguchi 89] and they were expressed using only essential case elements. Also, the target rules were English verb selection rules for four Japanese verbs "入る (hairu)", " 見える (mieru)" ," 見る (miru)", and " 取る (toru)". For these Japanese verbs, the number of real examples used was 95, 130, 385, and 33, respectively. The semantic hierarchy was the semantic hierarchy in ALT-J/E and the cross validation used was ten fold cross validation.

Case element values of artificial examples Usually, high level nodes on the Semantic hierarchy are used in the hand-made rules. On the other hand, example sentences have categories, which are leaves on semantic hierarchy, in the head noun. Thus, when the rules are converted into artificial examples (see Step-B in subsection 4.3), we can select
(1) *Leaf* that is a descendant of the category in the rules, or
(2) Any *descendant* of the category in the rules.
These two category selections will be evaluated in the following subsection.

Case element values of learned rules It is well known that ID3 employs "information gain" to select an attribute. In this paper, attributes are case elements. Usually, there are many case elements having the same information gain value. Thus, we will evaluate two types of learning :
(1) Upper node , on semantic hierarchy, is selected,
(2) Lower node , on semantic hierarchy, is selected.
Type (1) is called *Upper selection* and type (2) is called *Lower selection* in this evaluation.

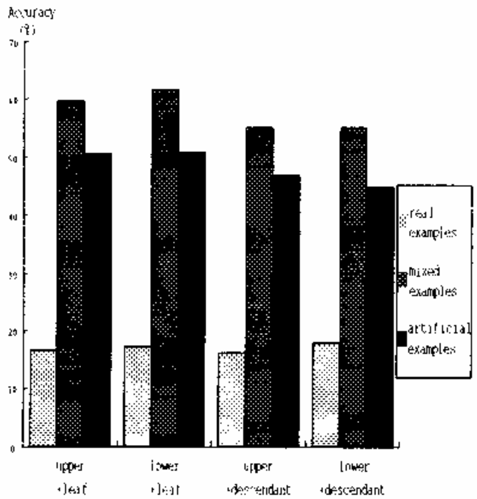

Figure 3: The experimental result of rules to J-verb 入る (hairu).

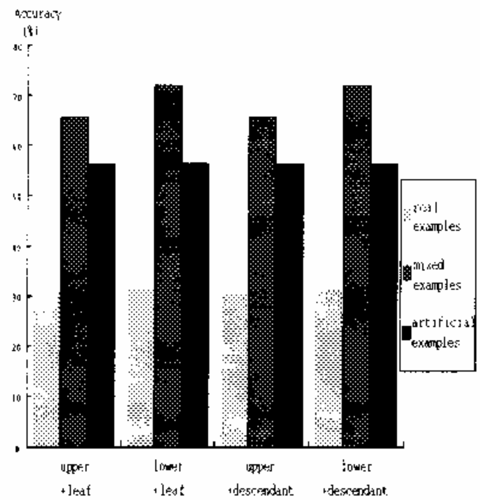

Figure 4: The experimental result of rules to J-verb 見える (mieru).

5.2 Result (1)

Figure 3 – Figure 6 show the experimental results. In these figures, *real examples* mean the accuracy of the rules constructed from only real examples and *artificial examples* mean the accuracy of the rules constructed from only artificial examples. Also, *mixed examples* are the accuracy of the rules constructed from both real examples and artificial examples. In the mixed examples case, candidate weighting values, in Step-ii of Subsection 4.2, are selected from the following candidates:

(weighting value of real examples, weighting value of artificial examples)
= (0.01,9.99), (0.1,9.9), (1,9), (2,8), (3,7), (4,8),
(5,5), (6,4),(7,3), (8,2), (9,1), (9.9,0.1), (9.99,0.01).

The accuracy of the mixed examples shown in Figure 3 – Figure 6 is the best accuracy achieved by these candidate weighting values. Please note that the sum of the weighting values of a real example and an artificial example equals ten. Other total values will be examined in the following subsection.

In these figures, *Upper (Lower) + Descendant* means that (1) the case element values of the learned rules are selected with upper (lower) node prior

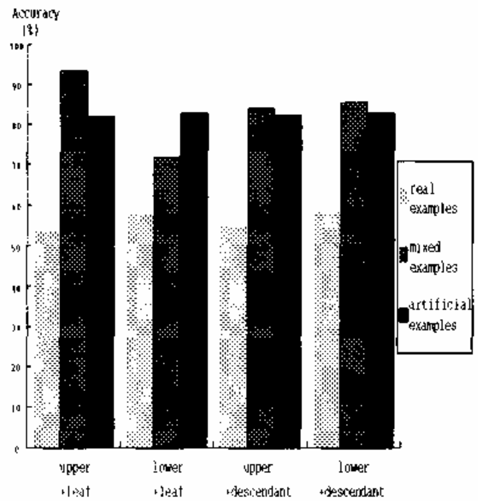

Figure 5: The experimental result of rules to J-verb 見る (miru).

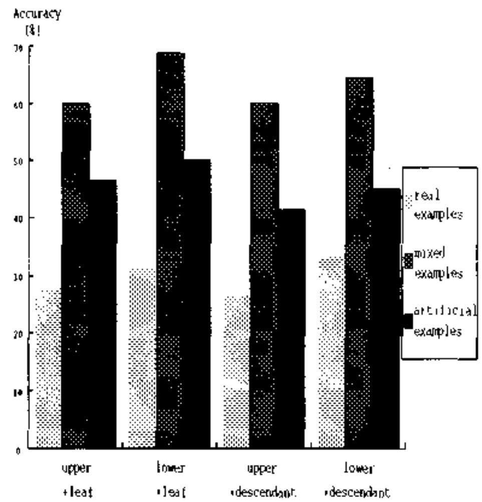

Figure 6: The experimental result of rules to J-verb 取る (toru).

in the internal learner, and (2) in the Artificial Example Generation step in Section 4.3, any descendant node is arbitrarily selected. Also, *Upper (Lower) + Leaf* means that (1) the case element values of the learned rules are selected with upper (lower) node priori, and (2) in the Artificial Example Generation step, the selection of a descendant is restricted to only a leaf.

As shown in Figure 3 – Figure 6, for every Japanese verb, not depending on the case element selection in artificial examples and restriction of descendant, English verb selection rules by our approach, i.e. "mixed examples", have better accuracy than either "real examples" or "artificial examples". Therefore, the proposed approach, using real examples and hand-made rules at the same time, overcomes for the shortage of real examples.

The best application of our approach is "lower + leaf" in the case of the three Japanese verbs "入る (hairu)", "見える (mieru)", and "取る (toru)", and "upper + leaf" in the case of the Japanese verb "見る (miru)". This difference among Japanese verbs depends on the number of real examples. That is, when many real examples can be gathered as for the Japanese verb "見る (miru)", English verb selection rules should be expressed using the upper semantic category. The reason is that the confidence level of the generalization using training examples increases when many real examples can be gathered.

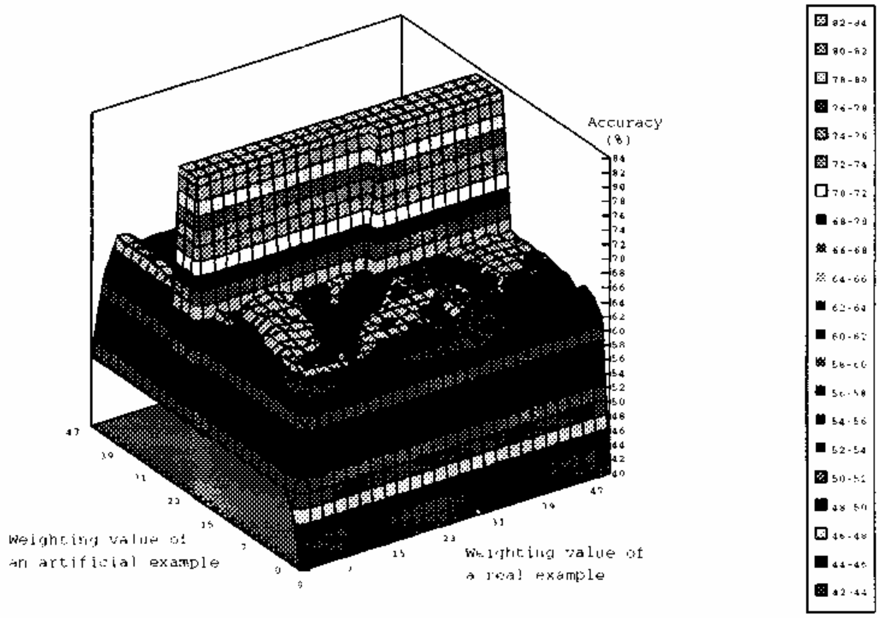

Figure 7: Result of cross validation.

For all Japanese verbs, the best descendant selection technique in step-B of Artificial Example Generation, is to restrict a descendant to a leaf.

The accuracy of the constructed English verb selection rules for the Japanese verb "見る (miru)" is much better than that for any other Japanese verb. This phenomenon is due to the fact that English verbs in the hand-made rules nearly equaled those in the real examples. Therefore, when constructing English verb selection rules using our approach, the hand-made rules should be prepared using English verbs as similar to those of the real examples as possible.

5.3 Result (2)

In the above subsection, the sum of the real example weighting value and the artificial example weighting value was always ten. Figure 7 shows the accuracy of the rules constructed from real examples and hand-made rules for the Japanese verb "取る (toru)". In this figure, the candidate weighting value of the real examples varies from 1 to 49 in steps of two. Also, the candidate weighting value of the artificial examples varies from 1 to 49 in steps of two. As a result, the pair of the above candidate weighting values form a lattice of odd integer pairs on $[1,49]\times[1,49]$.

Figure 7 shows that as the search space of candidate weighting values increases, higher accuracy can be achieved. Large search spaces, however, require longer running times. Because of the high learning speed of ID3[6], however, we can easily employ this technique if more accurate rules are necessary.

6 Conclusion

In this paper, a new method has been proposed that constructs English verb selection rules with high accuracy from hand-made English verb selection rules and sparse real examples. First, "artificial examples" are generated from hand-made English verb selection rules. Second, the artificial examples and the real examples are used as training examples for an internal learner. Finally, the internal learner outputs English verb selection rules offering improved translation quality. The main problem is determining the optimum weighting values. In this paper, the weighting values are fixed using cross validation.

In order to estimate the proposed method's performance, it was applied to hand-made English verb selection rules and real examples generated from a document, using Almuallim's learning algorithm as the internal learner. The English verb selection rules constructed by the proposed method have better accuracy than either those constructed from the only real examples or those that are hand-made.

References

[Almuallim 94a] Almuallim, H., Akiba, Y., Yamazaki, T., Yokoo, A., and Kaneda, S.,: "A Tool for the Acquisition of Japanese-English Machine Translation Rules using Inductive Learning Techniques", *Proc. CAIA-94*, pp.194-201, San Antonio, Texas, Mar. 1-4, 1994.

[6]The learning time of ID3 is nearly proportion to the number of training examples.

[Almuallim 94b] Almuallim, H., Akiba, Y., Yamazaki, T., and Kaneda, S.: "Induction of Japanese-English Translation Rules from ambiguous Examples and a Large Semantic Hierarchy", *Journal of Japanese Society for Artificial Intelligence,*, Vol.9, No.5, pp.730-740, Sept., 1994.

[Almuallim 94c] Almuallim, H., Akiba, Y., Yamazaki, T., and Kaneda, S.: "Two Methods for Learning ALT-J/E Translation Rules from Examples and A Semantic Hierarchy", *Proc. Coling-94*, pp.57-63, Kyoto, Aug., 5-9, 1994.

[Cultural Affairs 90] Agency for Cultural Affairs of Japan: *Dictionary of Basic Japanese Usage for Foreigners, (in Japanese),* Japanese Finance Publishing Agency, 1990.

[Horiguchi 89] Horiguchi,T.: *A handbook of illustrative sentences of English.* Nihon Tosho Library, 1989.

[Ikehara et al. 1989] Ikehara, S., Miyazaki, M.,Shirai,S. and Yokoo,A.: "An Approach to Machine Translation Method based on Constructive Process Theory", *Review of ECL*, Vol.37, No.1, pp.39-44, 1989.

[Ikehara et al. 1990] Ikehara,S.,Shirai, S., Yokoo, A. and Nakaiwa, H.: "Toward an MT System without Pre-Editing-Effects of New Methods in ALT-J/E", *Proc. MT Summit-3*, 1990.

[Keene 91] Donald Keene: *"Japanese-English Sentence Equivalents (Revised Version)"*, Asahi Press, 1991.

[Quinlan 86] Quinlan, J.R.: "Induction of Decision Trees", *Machine Learning*, 1(1), pp.81-106, 1986.

[Quinlan 92] Quinlan, J.R.: *C4.5 Programming for Machine Learning*, Morgan Kaufmann, San Mateo, California, 1992.