# Word Translation Based on Machine Learning Models Using Translation Memory and Corpora

**Kiyotaka Uchimoto†, Satoshi Sekine‡, Masaki Murata†, and Hitoshi Isahara†**

†Communications Research Laboratory
2-2-2, Hikari-dai, Seika-cho, Soraku-gun,
Kyoto, 619-0289 Japan
{uchimoto, murata, isahara}@crl.go.jp

‡New York University
715 Broadway, 7th floor
New York, NY 10003, USA
sekine@cs.nyu.edu

## Abstract

SENSEVAL-2 was held in Spring, 2001. It consisted of several tasks in various languages. In this paper, we describe our system used for one of these tasks: the Japanese translation task. With an accuracy of 63.4%, our system was the third best system in the contest among nine systems developed by seven groups.

## 1 Introduction

In the Japanese translation task, the senses of a word were defined in terms of the word's translations. Given an input sentence and a target word in the sentence, our system first estimates the similarity between the input sentence and parallel example sets called "Translation Memory". It then selects an appropriate translation of the target word by using the example set with the highest similarity. The similarity is calculated using dynamic programming and a machine learning model, which assesses the similarity based on the similarity of a string, words to the left and to the right of the target word in the input sentence, content words in the input sentence and their translations, and co-occurrence of content words in bilingual and monolingual corpora in English and Japanese.

## 2 Japanese Translation Task

In general, the definition of word senses depends on the goal of a task. The goal of the Japanese translation task is word selection in translation, where the target language is English. Therefore, word senses are defined as translations (translated words/phrases).

Before the contest, a Japanese-English parallel phrase/sentence set (Translation Memory, henceforth referred to as *TM*) was given to the participants as training data. In the TM, for each Japanese headword, there was a set of pairs of a Japanese expression including a headword and an English translation of the expression. We call these pairs *examples*. Some of the examples are shown in Figure 1.

```
<entry id="1" headword= "遠慮">
    <sense id="1-1">
        <jexpression> 母に遠慮する </jexpression>
        <eexpression>to feel constrained for one's
        mother</eexpression>
    </sense>
    <sense id="1-2">
        <jexpression> 母への遠慮 </jexpression>
        <eexpression>constraint toward one's
        mother</eexpression>
        <transmemo>UC</transmemo>
    </sense>
    <sense id="1-3">
        <jexpression> 献金を遠慮してもらう
        </jexpression>
        <eexpression>to request to refrain from
        donation</eexpression>
    </sense>
    .....
</entry>
```

Figure 1: Examples in TM.

In the formal test (contest), the participants were given a set of texts each of which was marked by a target word. For each target word, the participants were required to submit either a sense id of the example (the number assigned to each example in the TM), which can be used to translate the target word, or a translation of the target word. In the latter case, a translation of the word itself, a translation of a sequence of words including the target word, or a translation of the whole sentence could be submitted.

Answers were prepared for each target word in the formal test. The answers could consist of one or more sense id's in the TM, or of possible translations. The output of each system was evaluated in terms of accuracy, defined as a percentage of answers identified correctly by the system. An answer was judged to have been identified correctly when a sense id or a translation selected by the system was found in the answer.

## 3 Word Translation Model

Given an input sentence and a target word in the sentence, our model selects an appropriate translation of the target word or a sense id of examples appropriate for the translation of the target word by using examples with the highest similarity, estimated between the examples and the input sentence. In this paper, we call this model a *word translation model.* The source language is Japanese and the target language in translation is English. Henceforth we call a headword translation an *English headword.*

The similarity between an input sentence and examples is calculated by the following two methods:

1. A method based on the similarity of a string of characters (Method 1) : The similarity is defined as the amount of agreement between an input sentence and a Japanese example, expressed as a percentage.
2. A method based on machine learning models (Method 2) : The similarity is defined as the confidence or probability estimated by machine learning models. English headwords are used as classes (or categories) in machine learning models. Since the TM has examples with the same English headword, the similarity estimated by a model is the similarity between the input sentence and a set of examples.

A model is prepared for each Japanese headword. Given an input sentence, the similarity between the input sentence and each example is calculated by a model using Method 1. If the similarity is equal to or greater than a certain threshold, the model returns either the sense id of the example with the highest similarity or an English headword of the example. Otherwise, a model in Method 2 selects and returns an English headword.

The following sections describe the two methods in greater detail.

### 3.1 Method Based on the Similarity of A String of Characters (Method 1)

When an example with the highest similarity is found, it is given the highest priority, and either the sense id or the English headword of the example is selected as an output.

When calculating the agreement rate between an input sentence and an example, the rightmost word of the Japanese example is stemmed. In other words, when the rightmost word is a function word or a auxiliary verb such as "SURU (do)", it is eliminated. When the rightmost word is a predicate, its inflectional part is also eliminated. For example, the stemmed examples in Figure 1 are "母に遠慮", "母への遠慮", and "献金を遠慮", respectively. The agreement rate is calculated as a percentage of characters in the Japanese example that correspond to those in the input sentence. The correspondence is evaluated by comparing the Japanese example and the input sentence character by character. This can be done by using the UNIX command "diff" in a dynamic programming method. [1] The similarity is calculated by using the following equation.

$$Similarity = \frac{\left(\begin{array}{l}\text{the number of characters}\\ \text{corresponding to characters}\\ \text{in input sentence}\end{array}\right)}{\left(\begin{array}{l}\text{the number of characters in}\\ \text{stemmed Japanese example}\end{array}\right)} \quad (1)$$

When several examples with the highest similarity are found, the one having the longest Japanese example is selected except when the length of corresponding part is shorter than that of the Japanese headword.

However, it is unrealistic to expect that an example that is almost the same as the input sentence can be found because it is difficult to install all possible examples into the TM. So, when there is no example whose similarity is equal to or greater than the threshold, the method described in the next section is used.

### 3.2 Method Based on Machine Learning Models (Method 2) [2]

To select an appropriate example with the same usage as that of the input sentence, the similarity must be calculated by extracting the most important information from various conflicting sources of information related to the input sentence and examples. Since we want to avoid making complicated rules, we use machine learning models to calculate the similarity. Instead of all examples in the TM, English headwords are used as classes in machine learning models. Therefore, examples having the same English headword are put into the same class and are considered to have the same similarity.

[1] A description on how to use "diff" can be found in (Murata and Isahara, 2001).

[2] Work on using machine learning methods for the translation of tenses, aspects, and modalities can be found in (Murata et al., 2001a).

Classes identified by machine learning models are basically English headwords in TM, and they are detected manually. For example, English headwords of the examples in Figure 1 are "feel constrained", "constraint", and "refrain", respectively. When English headwords are verbs, they are represented by their basic forms. English words obtained when a Japanese headword is looked up in a Japanese-English dictionary are also used as classes.

For the training data, we use not only examples in the TM but also other data collected from bilingual dictionaries or a parallel corpus. The collected data consist of Japanese-English parallel phrases/sentences including both Japanese and English headwords, and they are used as complements of the training data.

For the machine learning models, we use SVM (Support Vector Machine), ME (Maximum Entropy), DL (Decision list), and SB (Simple Bayes). For each Japanese headword, the best model with the highest accuracy in 10-hold cross-validation on the training data is used for testing. The confidence of each class is estimated by probability distribution $p(a, b)$, where $b$ is a context in a set of contexts, $B$, and $a$ is a class in a set of classes, $A$. SVM is a classifier, and in this model, the confidence of each class cannot be represented by a probability distribution, but for the sake of convenience, we assign probability 1 to the most confident class estimated by SVM, and 0 to all other classes. The parameters in each model follow those used in (Murata et al., 2001b). Context $b$ is represented by a set of features, that is, information derivable from the training data. The features used in our experiments were as follows:

1. Morphological information
   The string, basic form, major and minor parts of speech, and inflection type on six morphemes, three morphemes to the left and three morphemes to the right of the target word in an input sentence.
2. Character n-gram
   Character n-grams in an input sentence. Each n-gram must include the target word.
3. Highest matching
   An English headword in the example that has the longest string matching that of the input sentence and its length are used as features.
4. Frequency of a content word and its translation candidates

   We define a set of examples including the same English headword as an example set. For each English headword, we define the following six example sets:

   **Example set 1** Japanese examples

   **Example set 2** English examples

   **Example set 3** Sentences similar to examples in Example set 1. They are collected from a Japanese monolingual corpus.

   **Example set 4** Sentences similar to examples in Example set 2. They are collected from an English monolingual corpus.

   **Example set 5** Union of Example sets 1 and 3

   **Example set 6** Union of Example sets 2 and 4

   For each example set, Japanese-English parallel phrases/sentences including both Japanese and English headwords are collected from bilingual dictionaries or parallel corpora, and are added to the example set.

   Sentences similar to a certain example are defined as sentences that include a substring of the example. The substring must include the headword of the example. In our model, we use sentences collected from a monolingual corpus because we want the model to reflect a real distribution of words, both headwords and words to the left and right of the headwords.

   As content words, we used nouns, verbs, adjectives, adverbs, and attributives, except headwords, in the input sentence. For each content word in an input sentence and its translation candidates, the frequencies in each example set were used as features. The translation candidates of a content word were obtained when the content word was looked up in a Japanese-English dictionary. Each feature is represented by a combination of an example set, a headword, and the frequency of content words in the example set. When we find that the total frequency of content words in an example set is $n$, we assume that every feature whose frequency is between 1 and $n$ is observed. For example, when the content word found in the given sentence is "mother", and it is found three times in the example set 1 for the headword "buy", the features "Example set 1 : buy : 1," "Example set 1 : buy : 2," and "Example set 1 : buy : 3" are assumed to be observed. By using these features, our model handles information about co-occurrence words of a headword in each corpus as a clue to translating the headword.

## 4 Experiment

### 4.1 Experimental conditions

The input and evaluation of the systems followed those of the Japanese translation task in SENSEVAL-2. A TM for 320 headwords was given to each participant in the middle of March, 2001. The average number of examples prepared for each headword was approximately 20. For the formal test, 40 target words (20 nouns and 20 verbs) were selected from the headwords. For each target word, 30 texts including the target words were prepared. The total number of the target words was 1,200.

As a bilingual dictionary, we used "EIJIRO" available at the web site of NIFTY, a network provider. As monolingual corpora, we used MAINICHI newspapers from 1991 to 2000, NIKKEI newspapers from 1995 to 1999, SANKEI newspapers from 1994 to 1999, and LDC data collected in 1994 and 1995, which include English newspaper articles for several years published by the Wall Street Journal, the Associated Press Writer, and the New York Times.

In the formal test, the threshold of similarity used in Method 1 was 1. JUMAN (Kurohashi and Nagao, 1999), a Japanese morphological analyzer, was used for morphological analysis in Method 2. As sentences similar to a certain example in Method 2, sentences that included a string obtained by stemming Japanese examples were extracted for Japanese examples, and sentences that included English headwords were extracted for English examples. As for the machine learning models, we could not select the most appropriate set of models by cross validation because not all learning processes could be finished by the deadline for submission. The models finally selected for the formal test were as follows:

- SVM : 23 words (12 nouns and 11 verbs)
- DL : 12 words (8 nouns and 4 verbs)
- SB : 5 words (5 verbs)

### 4.2 Experimental Results and Discussion

The accuracy obtained by our system in the formal test was 63.4% (761/1,200). The accuracy obtained by Method 1 and 2 were 91.0% (91/100) and 60.9% (670/1,100), respectively. Based on our results, we can draw the following conclusions:

- The system performance was related to the amount of training data per class in Method 2.
- The accuracy obtained for words whose English headwords were general words was not high even though there were more training data for these words than for other headwords for which the accuracy was high. We believe that this is due to the quality of automatically collected training data because general words appear in corpora quite frequently, and sometimes parallel sentences where Japanese and English headwords are not related to each other are collected. Therefore, we need to select automatically collected parallel sentences by aligning Japanese and English headwords.
- Method 1 improved the accuracy, especially for idiomatic expressions that rarely appeared in the training data. We applied Method 2 to the target words to which Method 1 was applied in the formal test, and achieved an even lower accuracy of 34.0%(34/100).
- The accuracy obtained by the SB model was low. We speculate that the SB model is not suitable for the feature sets used in the test.

## 5 Conclusion

This paper described our system used in SENSEVAL-2. Our model for word translation has the following characteristics: (1) It puts together examples having the same English headword into a set of examples, and selects a set of examples most similar to the input sentence by using machine learning models. (2) If an example that is almost the same as the input sentence is found, our model gives it the highest priority. (3) It automatically collects training data and information used for training from other language resources that are not only a bilingual corpus but also monolingual corpora of English and Japanese. We do not have to supervise anything except the detection of headword pairs in the examples.

## References


Sadao Kurohashi and Makoto Nagao, 1999. *Japanese Morphological Analysis System JUMAN Version 3.61.* Department of Informatics, Kyoto University.

Masaki Murata and Hitoshi Isahara. 2001. NLP using DIFF. In *IPSJ-WGNL NL144-18*, pages 127–134. (in Japanese).

Masaki Murata, Kiyotaka Uchimoto, Qing Ma, and Hitoshi Isahara. 2001a. Using a Support-Vector Machine for Japanese-to-English Translation of Tense, Aspect, and Modality. In *ACL Workshop on the Data-Driven Machine Translation.*

Masaki Murata, Masao Utiyama, Kiyotaka Uchimoto, Qing Ma, and Hitoshi Isahara. 2001b. Experiments on Word Sense Disambiguation Using Several Machine-leaning Methods. In *IEICE-WGNLC2001-2.* (in Japanese).