Collocational Analysis in Japanese Text Input

Masaki YAMASHINA Fumihiko OBASHI

NTT Electrical Communication Laboratories
1-2356 Take Yokosuka-shi Kanagawa-ken
238-03 JAPAN

Abstract

This paper proposes a new disambiguation method for Japanese text input. This method evaluates candidate sentences by measuring the number of Word Co-occurrence Patterns (WCP) included in the candidate sentences. An automatic WCP extraction method is also developed. An extraction experiment using the example sentences from dictionaries confirms that WCP can be collected automatically with an accuracy of 98.7% using syntactic analysis and some heuristic rules to eliminate erroneous extraction. Using this method, about 305,000 sets of WCP are collected. A co-occurrence pattern matrix with semantic categories is built based on these WCP. Using this matrix, the mean number of candidate sentences in Kana-to-Kanji translation is reduced to about 1/10 of those from existing morphological methods.

1.Introduction

For keyboard input of Japanese, Kana-to-kanji translation method [Kawada79] [Makino80] [Abe86] is the most popular technique. In this method, Kana input sentences are translated automatically into Kanji-Kana sentences. However, non-segmented Kana input is highly ambiguous, because of the segmentation ambiguities of Kana input into morphemes, and homonym ambiguities. Some research has been carried out mainly to overcome homonym ambiguity using a word usage dictionary [Makino80] and by using case grammar [Abe86].

A new technique named collocational analysis method, is proposed to overcome both ambiguities. This evaluates the certainty of candidate sentences by measuring the number of co-occurrence patterns between word pairs. It is used in addition to the usual morphological analysis. To realize this, it is essential to build a dictionary which can reflect Word Co-occurrence Patterns (WCP). In English processing research, there has been an attempt [Grishman86] to collect semi-automatically sublanguage selectional patterns. In Japanese processing research, there have been attempts [Shirai86] [Tanaka86] to collect combinations of words with this kind of relationship, either completely- or semi-automatically. These two attempts did not provide a dictionary for practical use.

A new method is proposed for building a dictionary which accumulates WCP. The first feature of this method is the collection of WCP from the common combination of two words having a dependency relationship in a sentence, because these common combinations will most likely reoccur in future texts. In this method, it is important to identify dependency relationships between word pairs, instead of identifying, the whole dependency structure of the sentence. For this purpose, Dependency Localization Analysis (DLA) is used. This identifies the word pairs having a definite dependency relationship using syntactic analysis and some heuristic rules.

This paper will first describe collocational analysis, a new concept in Kana-to-Kanji translation, then the compilation of WCP dictionary, next the translation algorithm and finally translation experimental results.

2. Concept of Collocational Analysis in Translation

Collocational analysis evaluates the correctness of a translated sentence by measuring the WCP within the sentence. The WCP data is accumulated in a 2-dimensional matrix, by information units indicating more restricted concepts than the words can indicate by themselves.

As previously mentioned there are two kinds of ambiguities in Kana-to-Kanji translation. In Fig.1, disambiguation process of homonyms is illustrated. ' 国歌 (a national anthem) and 演奏する (to play) ' and ' 国家 (a state) and 建設する (to build)' etc. are examples of WCP. If the simple Kana sequence ' こっかをえんそうする [kokkaoensousuru]' is input, the usual translation system will develop two possible candidate words ' 国歌 ' (a national anthem) and ' 国家 (a state)', for the partial Kana sequence of ' こっか [kokka]. The system will also develop uniquely the candidate word, ' 演奏する (to play) ' for ' えんそうする [ensousuru]'. These candidate words are obtained by table searching and morphological analysis. However, morphological analysis alone can't identify which one is correct for ' こっか [kokka]. Using collocational analysis, the WCP of '国家(a state)' and ' 演奏する (to play)' is found to be nil, while that of ' 国歌 (a national anthem)' and ' 演奏する (to play)' is found to be probable. Using WCP, ' 国歌を演奏する (to play a national anthem)' is selected as the final candidate sentence. If the Kana sequence ' こっかをけんせつする [kokkaokensetsusuru]' is input, ' 国家を建設する (to build a state)' is obtained in same manner.

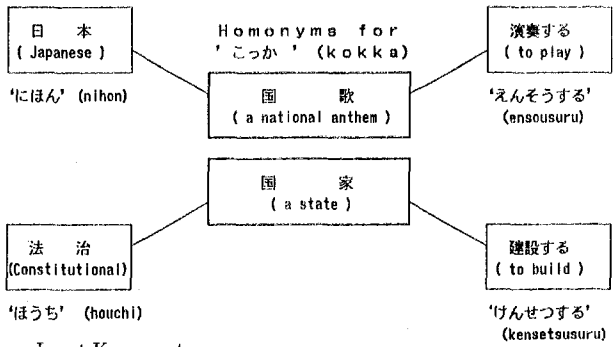


Fig. 1 Concept of collocational analysis

3. A WCP Dictionary

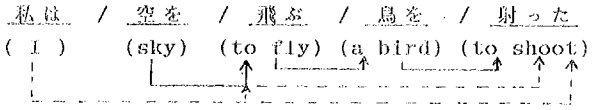
3.1 An Automatic Compilation Method

The new compilation method extracts from a sentence two word combinations which have a dependency relationship. This is illustrated with the sample sentence '私は空を飛ぶ鳥を射った (I shot a bird flying in the sky.)'.

Step.1

At first this method analyzes a sentence morphologically. In this example, the sentence is segmented into five Bunsetsu (Japanese grammatical units, like phrases) and the parts of speech of each word are obtained. '私 (I)', '鳥 (a bird)' and '空 (sky)' are nouns. '飛ぶ (to fly)' and '射った (to shoot)' are verbs. 'は (ha)' in the first Bunsetsu, 'を (o)' in the second one and in the fourth one are postpositional words. They determine the dependent attributes of nouns in dependency relationship.

ex.



(In English: I shot a bird flying in the sky.)

Step.2

The dependency relationship between words is analyzed using Japanese syntactic rules. In the extraction step, DLA is used. This process first finds out unique dependency relationships. "Unique relationship" means that a dependent has only one possible governor within the sentence. In this example, the relationships between '鳥を (a bird) and 射った (to shoot)' and '飛ぶ (to fly) and 鳥を (a bird)' are identified as unique.

Next, "ambiguous relationships" are processed. This relationship means that a dependent has several possible governors. In this case, the governor which can be identified as most likely by heuristic rules is located. This rule will only accept relationships where dependent and governor are adjacent, because this relationship has the highest possibility.

In this example, '空を(sky)' has two possible candidate governors, '飛ぶ (to fly)' and '射った (to shoot). In this case, because, '空を (sky) and 飛ぶ (to fly)' are adjacent, it is identified that '空を (sky)' is dependent and '飛ぶ (to fly)' is governor.

Next, '私は(I)' has also two possible candidate governors, '飛ぶ (to fly)' and '射った (to shoot)'. In this case, because, these two governors are not adjacent to the dependent, the dependency relationship between '私は(I)' and two candidate governors don't be identified for extraction.

Furthermore, some specific part-of-speech sequences which have many ambiguous dependency relationships are rejected for extraction. Following is an example of confusing part-of-speech sequence. In spite of similar syntactic style, '赤い (red)' in '赤い車の窓 (a red car's window)' modifys adjacent word '車 (a car)', while, '赤い(red)' in '赤い秋の花 (a red flower in fall)' modifys a word at end of sentence '花 (a flower)'. Thus, in case of this sequence, if a dependent and a governor are adjacent, the relationship between the modifying adjective and the modified noun is not identified.

ex.

modifying adj. etc. + noun + 'の'(postposition) + noun

(a red car's window) (red flower in fall)

3.2 Extraction Experiment

To provide a large volume of syntactically correct sentences, example sentences written in dictionaries [Ohno82] [Masuda83] were employed. This is because, these example sentences are a rich source of data indicating typical usage of each common word with short sentences and they are assumed to represent common usages within an extremely large amount of source data.

Five hundred example sentences were used to examine the accuracy of this automatic extraction method. 82% of sentences could be analyzed morphologically. As result, 718 sets of dependency relationship were extracted from these morphologically-analyzed sentences with an accuracy of 98.7%. The causes of erroneous extraction are mainly misidentification of part-of-speech and of compound words. The misidentification of dependency relationship was much less frequent.

Using this method, 305,000 sets of WCP were collected from 300,000 example sentences. In these WCP, about 45% of them are relationships between noun and verb or adjective with postpositional word, 21% are relationships between noun and noun with 'の (postpositional word)', and 26% are the nouns pairs constructing compound words.

3.3 Co-occurrence Pattern Matrix

With the aim of constructing a reliable WCP dictionary, the use of individual words, is impractical, because the dictionary becomes too large. Semantic categories are useful because, if word A and B are synonyms, they will have similar co-occurrence patterns to other words. This allows the WCP dictionary, described in semantic categories, to be greatly reduced in size. Scores of semantic categories were developed, however, it was found that the number of these categories was too small to accurately describe word relationships. Fortunately, there is a Japanese thesaurus [Ohno82] with 1,000 semantic categories. Based on the 305,000 sets of WCP data, a 2-dimensional matrix was developed which indicates co-occurrence patterns in semantic categories [ohno82].

Fig.2 shows an image of this matrix. In this matrix, word pairs which have same semantic categories have high co - occurrence possibility. The words included in the categories indicating 'action' and 'movement' etc. are the governor in a co-occurrence relationship with various words as their dependent.

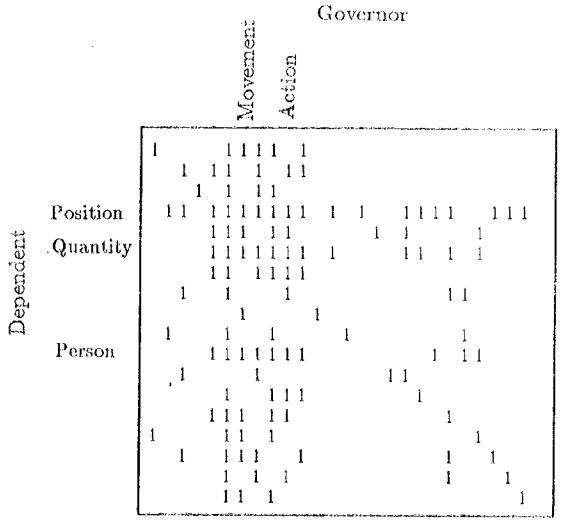


Fig. 2 An image of WCP matrix

4. Translation Algorithm

Fig.3 shows the translation process outline. First, table-searching is done for all segmentation possibilities to get each part-of-speech of segment. This is carried out referring to independent word dictionary (nouns, verbs, adjectives, etc. [65,000 words]), prefix and suffix dictionary [1085 words], dependent word dictionary (postpositions, auxiliary verbs, etc. [422 words]). Then, among the morpheme sequences constructed with each segment, the grammatically possible sequences are selected.

Next, the candidate sentences with the least number of Bunsetsu are selected [Yoshimura83]. Furthermore, among these selected sentences, those which have the least number of words are selected. In this process, a heuristic rule is used to prevent morpheme sequence mis-selection. This rule rejects the combinations of nouns constructing a compound word, if the usage frequency of either nouns is very low.

ex.

Input Kana sequence: かんけいのなか [kankeinonaka]

× 関係 (noun) 野中 (noun, freq. : very low)
(a relation) (in a field)

○ 関係 (noun) の (postposition) 中 (noun, freq. : high)
(a relation) (among)

Secondly, the co-occurrence pattern matrix is utilized in order to determine the number of WCP within each candidate sentence. The counting operation is carried out only on adjacent Bunsetsu, because, in most cases, relationships are between adjacent Bunsetsu and determining extended relationships would prove to be too time-consuming.

Finally, the candidate sentence with the maximum WCP number is chosen as the prime candidate. To prevent mistaken deletion of prime candidates caused by word pairs which rarely co-occur, following rule is used. If the usage frequency of either word in WCP is low, the candidate sentences of which WCP number is less one than maximum number, are also identified as prime candidates. In following example, both are identified as prime candidates.

ex

Input Kana sequence: ぶんしょうのこうせい
[bunshounokousei]

文章 (a sentence) の 校正 (freq. : low) (proofreading)
WCP

文章 (a sentence) の 構成 (freq. : high) (composition)
not WCP

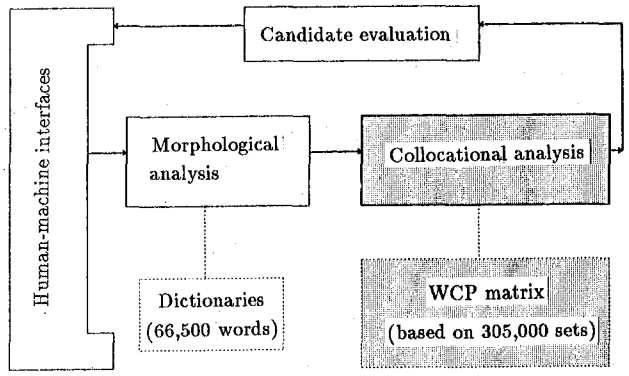


Fig. 3 Translation process outline

5. Translation Experimental Result

About four hundred test sentences were used to evaluate the accuracy of collocational analysis. The mean number of candidate sentences was 62.6, selected by considering least number of Bunsetsu. Error ratio for this was 1.7%. Error ratio means the proportion of correct Bunsetsu missed by the selecting operation in each process to total number of all Bunsetsu. The mean number of candidate sentences selected by least number of words was 16.1 with an error ratio of 0.8%. Finally, the number of candidate sentences selected by collocational analysis method was further reduced to 6.4 with an error ratio of 1.6%.

Furthermore, translation accuracy of the practical translation algorithm based on the above description was examined using 10 leading articles in news papers(about 14,000 characters). This practical algorithm was modified for processing proper nouns, numerals and symbols, and to save memory. It was confirmed that the translation accuracy evaluated by character unit of this method was higher than 95%.

6. Conclusion

A method for disambiguation based on collocational analysis of non-segmented Kana-to-Kanji translation has been developed. To realize this, an automatic WCP dictionary compilation method has also been developed. In an extraction experiment using example sentences from dictionaries, it was confirmed that WCP can be collected automatically with a 98.7% accuracy using syntactic analysis and some heuristic rules to eliminate errors. Using this method, about 305,000 sets of WCP were collected. The co-occurrence pattern matrix was built based on these WCP and used in translation experiments.

Experimental results show that the mean number of candidate sentences is reduced to about 1/10 of those from existing morphological methods and that a translation accuracy of 95% can be achieved. The collocational analysis method can also be applied to Japanese text input by speech recognition.

Reference

Abe, M., et al.(1986), "A Kana-Kanji Translation System for Non-Segmented Input Sentences Based on Syntactic and Semantic Analysis", Proceeding of COLING86, 280-285

Grishman, R., et al.(1986), "Discovery Procedures for sublanguage Selectional Patterns", Computational Linguistics, vol.12, no.3,205-215

Kawada, t., et al.(1979), "Japanese Word Processor JW-10", Proceeding of COMPCOM'79 fall, 238-242

Makino, H., et al. (1980), "An Automatic Translation system of Non-segmented Kana Sentences into Kanji-Kana sentences", Proceeding of COLING, 295-302

Masuda, K., et al. (1983), "Kenkyusya's New Japanese-English Dictionary", Kenkyusya, Tokyo

Ohno, S., et al.(1982), "New Synonyms Dictionary" (in Japanese), Kadokawa-syoten, Tokyo

Shirai, K., et al.(1986). "Linguistic Knowledge Extraction from Real Language Behavior", Proceeding of COLING86, 253-255

Tanaka, Y., et al.(1986), "Acquisition of Knowledge Data by Analyzing Natural Language", Proceeding of COLING86, 448-450

Yoshimura, K., et al.(1983), "Morphological Analysis of Nonmarked-off Japanese Sentences by the least BUNSETSU's Number Method", Johoshori, Vol.24, No.1, 44-46