

Collocations as Word Co-Occurrence Restriction Data

-An Application to Japanese Word Processor -

K osho Shudo*, M asahito Takahashi*, Y asuo K oyam a**, K enji Y oshimura*

*Faculty of Engineering, Fukuoka University *Fukuoka 814-0180, Japan

**aisoft.co. **2-1-27, Chuou, M atsumoto, Nagano 390-0811, Japan

*shudo@ tl.fukuoka-u.ac.jp, *takahashi@ helio.tl.fukuoka-u.ac.jp,

**koyama@ aissoft.co.jp, *yoshimura@ tl.fukuoka-u.ac.jp

Abstract

Collocations, the combination of specific words are quite useful linguistic resources for NLP in general. The purpose of this paper is to show their usefulness, exemplifying an application to Kanji character decision processes for Japanese word processors. Unlike recent trials of automatic extraction, our collocations were collected manually through many years of intensive investigation of corpus. Our collection procedure consists of (1) finding a proper combination of words in a corpus and (2) recollecting similar combinations of words, incited by it. This procedure, which depends on human judgment and the enrichment of data by association, is effective for remedying the sparseness of data problem, although the arbitrariness of human judgment is inevitable. Approximately seventy two thousand and four hundred collocations were used as word co-occurrence restriction data for deciding Kanji characters in the processing of Japanese word processors. Experiments have shown that the collocation data yield 8.9% higher fraction of Kanji-to-Kanji character conversion accuracy than the system which uses no collocation data and 7.0% higher, than a commercial word processor software of average performance.

1. Introduction

The meanings of words in a natural sentence are mutually bound. The most crucial problem for any kind of NLP is to disambiguate the meanings of each word in the sentence with clues from word co-occurrence. It is, however, well known that it's very difficult to construct a rule system which disambiguates them effectively for unrestricted or less-restricted natural sentences. Authors have pointed out the importance of lexicon-based linguistic resources for prescribing the co-occurrence of words such as collocations, idioms etc., as well as rule-based resources, e.g., the set of case patterns combined with a thesaurus or a set of semantic features. (Shudo, 1989)

In particular, collocations, which are the combination of specific words are quite important linguistic resources for NLP in general. The purpose of this paper is to show their usefulness, exemplifying their application to Kanji character decision processes for Japanese word processors.

2. Collocation Data

Our collocations were collected manually through many years of intensive investigation of corpus. The extraction procedure consists of (1) finding a proper combination of words in a corpus and (2) recollecting similar combinations of words, being incited by it. This procedure, which depends on human judgment and enrichment of data by association, seems effective both to remedy the data sparseness problem and to reduce the noise of the data, although the arbitrariness of human judgment to some extent is inevitable. No automated, stochastic method for the current state of the art assures the necessity and sufficiency of data, in spite of recent ambitious trials. (Shinnoh, 1995, Ikehara, 1996)

The collocations which we collected are word strings whose specific component word implies strongly the occurrence of the rest. For example, "ぐっすり眠る gussuri nemuru(sleep soundly)" is collected as a

collocation since the occurrence of "ぐっすり gussuri(soundly)" strongly implies the occurrence of "眠る nemuru(sleep)" in the close position, namely the conditional probability $p(\text{"眠る sleep"} | \text{"ぐっすり soundly"})$ of occurrence seems significantly larger than the case for ordinary verb-adverb pairs. Presumably several million sentences or phrases in printed articles such as newspapers, school textbooks, journals and dictionaries which were investigated. We believe that in spite of some arbitrariness, the ability of the educated adult to recollect expressions of his/her native language and to estimate their frequency is more superior than any current automated system based on n-gram models et al. Most Japanese idioms and proverbs are included in the data. The collocations are classified into two classes by their syntactic functions; one is a class of functional collocations which work like functional words. The other is a class of conceptual collocations which have conceptual contents like nouns, verbs, adverbs, adjectives, adjective verbs.

3. Functional Collocations

We have two kinds of functional words; one is the particle (postpositional) which is used to mark a certain relationship between concepts in the sentence, and the other is the auxiliary verb which is used to give the sentence tense, aspect, mood, mode, speaker's judgments, etc. Accordingly, functional collocations are classified into two types, as well; one is relational collocation and the other is auxiliary predicative collocations. (Shudo, 1979) The former is exemplified by expressions in English such as "in order to", "based on", "according to", "in proportion to", etc. and the latter, by "is able to", "have to", "is obliged to", etc.

3.1 Relational Collocations

Approximately one thousand relational collocations were collected. Table 1 shows examples of them. "/" in the table shows the word boundary in the collocation.

について ni/tuite (about),
によって,にあける,に基づいて,
のように,でさえも,などといった,
だけでなく,にも/かかわらず,とは言え,
に対し,ばかりでなく,の通りに

Table 1: Examples of the relational collocation

なければ/ならない nakereba/naranai (must),
かもしれない,てりる,たほうが/よい,
べきである,に違いない,おそれがある,
すらない,とは限らない,
ずにはいられない,には及ばない

Table 2: Examples of the auxiliary predictive collocation

3.2 Auxiliary Predictive Collocation

Approximately one thousand four hundred auxiliary predictive collocations were collected. Table 2 shows some examples of them.

4. Conceptual Collocation

Approximately seventy thousand conceptual collocations were collected. Table 3. shows examples of conceptual collocations.

5. Application to Kana-to-Kanji Conversion for Japanese word Processor

5.1. Japanese Word Processor

A Japanese word processor receives words, phrases or sentences written exclusively in Kana (phonetic) characters through keyboard strokes and outputs their equivalents written in Kana and Kanji (ideographic) characters. It is a kind of Machine Translation system which converts source Kana character strings into targets of Kana and Kanji mixed strings. Usually, these strings have no space between words. Input strings are analyzed morphologically and segmented into words (or morphemes). Then, words which should be converted into Kanji strings and their equivalent Kanji strings are chosen. The major issue for this technology is concerned with raising the accuracy of the segmentation into words and of the selection of Kanji among many homophonic candidates. Besides the semantic framework, e.g., thesaurus-based or semantic feature-based restrictions on word co-occurrences, case frame, neural net, etc., which have been adopted in recent works(Oshima, 1986; Kobayashi, 1987; Yamamoto, 1992), we attempted to raise the accuracy of Kanji selection by adopting surface level resources, i.e., collocation data explained in Chapter 2-4.

5.2 Experiment

We adopted the minimum cost method(Yoshimura, 1987) combined with Viterbi algorithm(Viterbi, 1967) in the disambiguation of segmentations. Figure 1 illustrates the segmentation and conversion process of input string

kind	Example
Nominal	赤の / 他人 akano/tanin (a perfect stranger), 目の毒, 大掴み, 貰い泣き, ラッパ飲み, 悪業の報い, 環境の汚染, 頬に汗, お手数をおかけ
Verbal	おつりが来る otsuriga/kuru (be enough to make change), 相異なる, 汗水/垂らす, 大人びる, 一息/入れる, 堀抜ける, 心が沈む, 気を吐く, 座を占める, 体をあけ
Adjectival	裏悲しい uraganashii (mournful), 青臭い, えげつない, 途方もない, 態度がでかい, 気が重い
Adjective-verbal	御機嫌斜め gokigen/hanam e (in a bad mood), 誇らしげ, 懐が暖か futokoroga/ atataka (be rich), 愛情が細やか
Adverbial	案の定 anno/jyu (as expected), いつになく, 目を輝かせて, 先を争って
Adnominal	悪しき ashiki (bad), あじな
Four-Kanji-compound	我田引水 gaden'insui (every miller draws water to his own mill), 不惜身命
~suru	公にする ooyakeni/suru (make public), 相手にする
Others	年端も/いかぬ toshihamo/ikanu (young), 如何せん, 間尺に合わぬ, 心胆寒からしむ

Table 3: Examples of conceptual collocations

"korehanegatotem ofukai (This has a very deep cause.)". A line denotes possible concatenation of candidate words (unitary expression -bunsetsu). The number attached to each word candidate means the partial minimum cost of the string of words which begins at the top of the input string and ends with the word. A collocation, "根が深い nega/fukai" interrupted by an adverb "とても totemo" and succession of adverb-adjective, i.e., "とても ~深い" causes the addition of cost -3 and -1, respectively, to the basic cost, +8 of the string "これは根がとても深い". Thus, the minimum cost (+4) of "これは根がとても深い" is obtained as the result.

5.3 Prototype System A

We first developed a prototype Kana-to-Kanji conversion system named system A which is equipped with no collocation data but with only an ordinary word dictionary.

5.4 System B, C and D

We next reinforced System A to obtain system B by adding collocation data. System C is a slightly modified commercial system, i.e. W X G Ver2.05 (made by a.i. soft co.) adjusted to our experimental framework. System D is system C, additionally equipped with the collocation data.

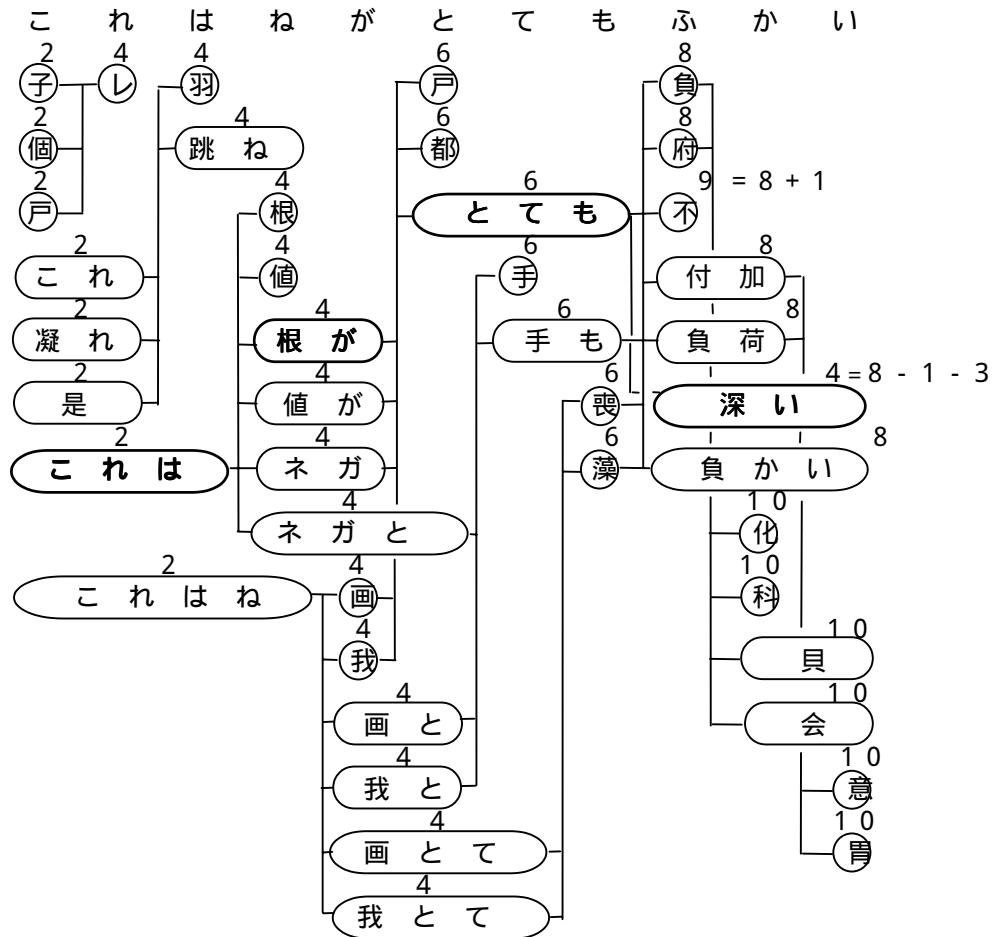


Figure 1: An example of processing

5.5 Input Data for Evaluation

We prepared approximately twenty three thousand pairs of input Kana strings and their equivalent Kana-Kanji mixed strings to be the expected output for our experiment. The latter includes additional information about each word boundary with the acceptable tolerance. The average number of Kana for each input string is approximately twenty eight.

5.6 Experimental Results

The experimental system performs the morphological analysis along with the cost calculation, selects the least cost segmentation and Kanji candidate as the conversion result and then checks whether it matches the acceptable answer given with the input string. The system gives the statistical data after processing all input strings, calculating inputs successfully segmented and converted. The major results are given in Table 4.

Comparing System A with B, we conclude that the introduction of approximately 72,400 collocations cause an 8.9% rise of accuracy rate. In addition, the collocation data provide a 7.0% higher accuracy rate for System D than System C.

5.7 Input Length vs. Accuracy

	System A	System B	System C	System D
Conversion	10,978 /22,923	13,013 /22,923	12,424 /22,923	14,027 /22,923
	47.9%	56.8%	54.2%	61.2%
Segmentation	18,424 /22,923	19,555 /22,923	18,620 /22,923	19,682 /22,923
	80.4%	85.3%	81.2%	85.9%

Table 4: Results of the experiment

Figure 2 shows the relationship between the length of input string and the accuracy of the conversion of System D. More than 520,000 different segmentations are mathematically possible in situations where the length of input is twenty Kana and the minimum and maximum length of a segment is one and eighteen, respectively. On the other hand, Figure 2 shows that the segmentation accuracy rate for a twenty Kana input is 90.7%. Our figures show that combinatorial explosion of ambiguity is well prevented by the linguistic and technological framework of our system, despite longer inputs and greater demands on the system.

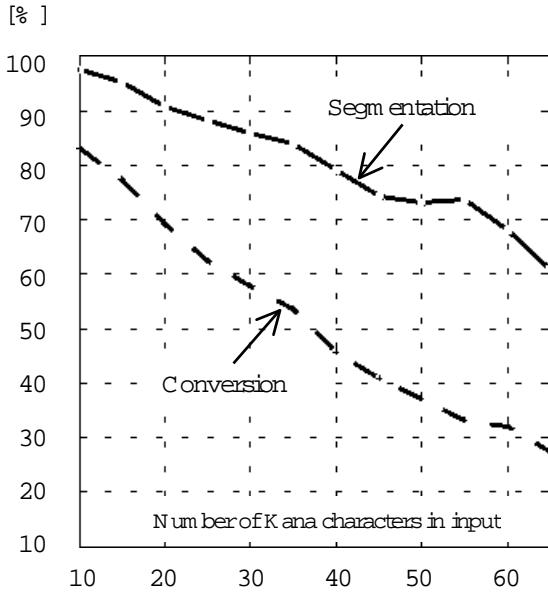


Figure 2: Accuracy vs. length of input (System D)

6. Concluding Remarks

Techniques using extensive surface linguistic resources are quite important for the future evolution of NLP. The prospect of the availability of large scale collocation data for the reduction of ambiguity raised for various kinds of NLP is presented in this paper, showing an experimental example of their application to the Japanese word processor. In fact, they work as an effective restriction for word co-occurrence whereas the word-class level prescription tends towards various failures.

Collocations have been collected manually from more than ten years by the authors in order to improve, not only a high precision word processor, but also more general Japanese language processing systems. Many resources, e.g. school textbooks, newspapers, novels, journals, dictionaries, etc., were referred to by workers, who then proceeded to judge likely candidates for the collocation. This process is based on our contention that linguistic expressions are extracted from more accurately and from more extensively through humans than by computer program, despite certain arbitrariness inherent in human judgement and their selection of collocations. The sparseness problem of expression data is crucial in the automatic and stochastic approach. On the other hand, the brain of the educated adult is quite powerful in storing and retrieving expressions of his/her mother tongue by association.

We believe the next attractive and promising field of application for our data will be large vocabulary continuous speech recognition.

References

Shudo K. et al. (1979). A Structural Model of Bunsetsu for Machine Processing of Japanese. In Trans. IECE, 62-D-12, (in Japanese).

Shudo K. (1989). Japanese Collocations. Technical Report of Grant-in-Aid for Scientific Research, No.63101005, MESSC, (in Japanese).

Shinnou H & Isahara, H. (1995). Automatic Acquisition of Idioms on Lexical Peculiarity. In Trans. of IPSJ, 36-8,

(in Japanese).

Ikehara S. et al. (1996). A Statistical Method for Extracting Uninterrupted and Interrupted Collocations from Very Large Corpora. In Proc. of 16th Internat. Conf. on Computational Linguistics (COLING 96).

Oshima Y. et al. (1986). A Disambiguation Method in Kana-to-Kanji Conversion Using Case Frame Grammar. In Trans. of IPSJ, 27-7, (in Japanese).

Kobayashi T. et al. (1987). Realization of Kana-to-Kanji Conversion Using Neural Networks. In Toshiba Review, 47-11, (in Japanese).

Yamamoto K. et al. (1992). Kana-to-Kanji Conversion Using Co-occurrence Groups. In Proc. of 44th Conf. of IPSJ, (in Japanese).

Yoshimura K. et al. (1987). Morphological Analysis of Japanese Sentences using the Least Cost Method", In IPSJ SIG NL-60, (in Japanese).

Viterbi A. J. (1967). Error Bounds for Convolutional Codes and an Asymptotically Optimal Decoding Algorithm. In IEEE Trans. on Information Theory, 13.

Church K. W. et al. (1990). Word Association Norms, Mutual Information, and Lexicography. In Computational Linguistics, 16.