Large Scale Collocation Data and Their Application to Japanese Word Processor Technology

Yasuo Koyama, Masako Yasutake, Kenji Yoshimura and Kosho Shudo
Institute for Information and Control Systems, Fukuoka University
Nanakuma, Fukuoka, 814-0180 Japan
koyama@aisoft.co.jp, yasutake@helio.tl.fukuoka-u.ac.jp, yosimura@tlsun.tl.fukuoka-u.ac.jp,
shudo@tlsun.tl.fukuoka-u.ac.jp

abstract

Word processors or computers used in Japan employ Japanese input method through keyboard stroke combined with Kana (phonetic) character to Kanji (ideographic, Chinese) character conversion technology. The key factor of Kana-to-Kanji conversion technology is how to raise the accuracy of the conversion through the homophone processing, since we have so many homophonic Kanjis. In this paper, we report the results of our Kana-to-Kanji conversion experiments which embody the homophone processing based on large scale collocation data. It is shown that approximately 135,000 collocations yield 9.1 % raise of the conversion accuracy compared with the prototype system which has no collocation data.

1. Introduction

Word processors or computers used in Japan ordinarily employ Japanese input method through keyboard stroke combined with Kana (phonetic) to Kanji (ideographic, Chinese) character conversion technology. The Kana-to-Kanji conversion is performed by the morphological analysis on the input Kana string with no space between words. Word- or phrase-segmentation is carried out by the analysis to identify the substring of the input which has to be converted from Kana to Kanji. Kana-Kanji mixed string, which is the ordinary form of Japanese written text, is obtained as the final result. The major issue of this technology lies in raising the accuracy of the segmentation and the homophone processing to select the correct Kanji among many homophonic candidates.

The conventional methodology for processing homophones have used the function that gives the priority to the word which was used lastly or to the high frequency word. In fact, however, this method sometimes tends to cause inadequate conversion due to the lack of consideration of the semantic consistency of the word concurrence. While it is difficult to employ the syntactic or semantic processing in earnest for the word processor from the cost vs. performance viewpoints, for example, the following trials to improve the conversion accuracy have been reported: Employing the case-frame to check the semantic consistency of combination of words [Oshima, Y. et al.,1986]. Employing the neural network to describe the consistency of the concurrence of words [Kobayashi,T. et al.,1992], Making a concurrence dictionary for the specific topic or field, and giving the priority to the word which is in the dictionary when the topic is identified [Yamamoto, K. et al., 1992]. In any of these studies, however, many problems are left unsolved in realizing its practical system.

Besides these semantic or quasi-semantic gadgets, we think it much more practical and effective to use surface level resources, namely, to use extensively the collocation. But how many collocations contribute to the accuracy of Kana-to-Kanji conversion is not known yet.

In this paper, we present some results of our experiments of Kana-to-Kanji conversion, focusing on the usage of large scale collocation data. In chapter 2, descriptions of the collocations used in our system and their classification are given. In chapter 3, the technological framework of our Kana-to-Kanji conversion systems is outlined. In chapter 4, the method and the results of the experiments are given along with some discussions. In chapter 5, concluding remarks are given.

2. Collocation Data

Unlike the recent works on the automatic extraction of collocations from corpus [Church, K. W, et al, 1990, Ikehara, S. et al, 1996, etc.], our data have been collected manually through the intensive investigation of various texts, spending years on it. This is because no stochastic framework assures the

accuracy of the extraction, namely the necessity and sufficiency of the data set. The collocations which are used in our Kana-to-Kanji conversion system consist of two kinds: (1) idiomatic expressions, whose meanings seem to be difficult to compose from the typical meaning of the individual component words [Shudo, K. et al., 1988]. (2) stereotypical expressions in which the concurrence of component words is seen in the texts with high frequency. The collocations are also classified into two classes by a grammatical criterion: one is a class of **functional** collocations, which work as functional words such as particles (postpositionals) or auxiliary verbs, the other is a class of **conceptual** collocations which work as nouns, verbs, adjectives, adverbs, etc. The latter is further classified into two kinds: **uninterruptible** collocations, whose concurrence relationship of words are so strong that they can be dealt with as single words, and **interruptible** collocations, which are occasionally used separately.
In the following, the parenthesized number is the number of expressions adopted in the system.

2.1 Functional Collocations (2,174)

We call expressions which work like a particle **relational** collocation and expressions which work like an auxiliary verb at the end of the predicate **auxiliary predicative** collocation [Shudo, K. et al. ,1980].

relational collocations (760)
ex. に/ついて
ni/tuite (about)

auxiliary predicative collocations (1,414)
ex. なければ/ならない
nakereba/naranai (must)

2.2 Uninterruptible Conceptual Collocations (54,290)

four-Kanji-compound (2,231)
ex. 我田引水
gadeninsui
(every miller draws water to his own mill)

adverb + particle type (3,089)
ex. あたふたと
atafutato (disconcertedly)

adverb + suru type (1,043)
ex. あくせくする
akusekusuru (toil and moil)

noun type (21,128)
ex. 赤の/他人
akano/tanin (perfect stranger)

verb type (13,225)
ex. おつりが/来る
otsuriga/kuru
(be enough to make the change)

adjective type (2,394)
ex. 裏悲しい
uraganashii (mournful)

adjective verb type (397)
ex. ご機嫌/斜め
gokigen/naname (in a bad mood)

adverb and other type (8,185)
ex. 目に/見えて
meni/miete (remarkably)

proverb type (2,598)
ex. 老いては/子に/従え
oiteha/koni/shitagae
(when old, obey your children)

2.3 Interruptible Conceptual Collocations (78,251)

noun type (7,627)
ex. 悪行の/報い
akugyouno/mukui (fruit of an evil deed)

verb type (64,087)
ex. 後ろ髪を/引かれる
ushirogamiwo/hikareru
(feel as if one's heart were left behind)

adjective type (3,617)
ex. 態度が/大きい
taidoga/ookii (act in a lordly manner)

adjective verb type (2,018)
ex. 役者が/上
yakushaga/ue (be more able)

others (902)
ex. 後に/引けぬ
atoni/hikenu (can not give up)

3. Kana-to-Kanji Conversion Systems

We developed four different Kana-to-Kanji conversion systems, phasing in the collocation data described in **2**. The technological framework of the system is based on **extended bunsetsu** (**e-bunsetsu**) model [Shudo, K. et al., 1980] for the unit of the segmentation of the input Kana string, and on **minimum cost method** [Yoshimura,K. et al., 1987] combined with Viterbi's algorithm [Viterbi, A,, J., 1967] for the reduction of the ambiguity of the segmentation.
A **bunsetsu** is the basic postpositional or predicative

phrase which composes Japanese sentences, and an **e-bunsetsu**, which is a natural extension of the bunsetsu, is defined roughly as follows:

<e-bunsetsu>::= <prefix>* <conceptual word | uninterruptible conceptual collocation> <suffix>* <functional word | functional collocation>*

The e-bunsetsu which includes no collocation is the bunsetsu. More refined rules are used in the actual segmentation process. The interruptible conceptual collocation is not treated as a single unit but as a string of bunsetsus in the segmentation process.

Each collocation in the dictionary which is composed of multiple number of bunsetsus is marked with the boundary between bunsetsus. The system first tries to segment the input Kana string into e-bunsetsus. Every possible segmentation is evaluated by its cost. A segmentation which is assigned the least cost is chosen as the solution.

The boundary between e-bunsetsus in examples in this paper is denoted by "/".

ex. two results of e-bunsetsu-segmentation:

人は/気が利くに越した事は有りません
hitoha/kigakikunikositakotohaarimasen
(there is nothing like being watchful)
人は/気が/利くに/越した/事は/有りません
hitoha/kiga/kikuni/kosita/kotoha/arimasen

In the above examples, 気が/利く *kiga/kiku*: is uninterruptible conceptual collocation and に/越した/事は/有りません *ni/kosita/kotoha/arimasen*: is a functional collocation. In the first example, these collocations are dealt with a single words. The second example shows the conventional bunsetsu-segmentation.

The cost for the segmentation candidate is the sum of three partial costs: b-cost, c-cost and d-cost shown below.

(1)a segment cost is assigned to each segment. Sum of segment costs of all segments is the basic cost (b-cost) of a segmentation candidate. By this, the collocation tends to have priority over the ordinary word. The standard and initial value of each segment cost is 2, and it is increased by 1 for each occurrence of the prefix, suffix, etc. in the segment.

(2)a concatenation cost (c-cost) is assigned to specific e-bunsetsu boundaries to revise the b-cost. The concatenation, such as adnominal-noun, adverb-verb, noun-noun, etc. is paid a bonus , namely a negative cost, -1.

(3)a dependency cost (d-cost), which has a negative value, is assigned to the strong dependency relationship between conceptual words in the candidate, representing the consistency of concurrence of conceptual words. By this, the segmentation containing the interrupted conceptual collocation tends to have priority. The value of a d-cost varies from –3 to –1, depending on the strength of the concurrence. The interruptible conceptual collocation is given the biggest bonus i.e.-3.

The reduction of the homophonic ambiguity, which limits Kanji candidates, is carried out in the course of the segmentation and its evaluation by the cost.

3.1 Prototype System A

We first developed a prototype Kana-to-Kanji conversion system which we call System A, revising Kana-to-Kanji conversion software on the market, WXG Ver2.05 for PC.

System A has no collocation data but conventional lexical resources, namely functional words (1,010) and conceptual words (131,661).

3.2 System B, C and D

We reinforced System A to obtain System B, C and D by phasing in the following collocational resources. System B is System A equipped additionally with functional collocations (2,174) and uninterruptible conceptual collocations except for four-Kanji-compound and proverb type collocations (49,461). System C is System B equipped additionally with four-Kanji-compound (2,231) and proverb type collocations (2,598). Further, System D is System C equipped additionally with interruptible conceptual collocations (78,251).

4. Experiments

4.1 Text Data for Evaluation

Prior to the experiments of Kana-to-Kanji conversion, we prepared a large volume of text data by hand which is formally a set of triples whose first component **a** is a Kana string (a sentence) with no space, The second component **b** is the correct segmentation result of **a**, indicating each boundary between bunsetsus with "/" or ".". "/" and "." means obligatory and optional boundary, respectively. The third component **c** is the correct conversion result of **a**, which is a Kana-Kanji mixed string.

ex. { **a**: にわにばらがさいている
niwanibaragasaiteiru

(roses are in bloom in a garden)
b: にわに/ばらが/さいて.いる
niwani/baraga/saite.iru
c: 庭に/バラが/咲いて.いる }

The introduction of the optional boundary assures the flexible evaluation. For example, each of 咲いて/いる *saite/iru* (be in bloom) and 咲いている *saiteiru* is accepted as a correct result. The data file is divided into two sub-files, f1 and f2, depending on the number of bunsetsus in the Kana string **a**. f1 has 10,733 triples, whose **a** has less than five bunsetsus and f2 has 12,192 triples, whose **a** has more than four bunsetsus.

4.2 Method of Evaluation

Each **a** in the text data is fed to the conversion system. The system outputs two forms of the least cost result: **b'**, Kana string segmented to bunsetsus by "/" , and **c'**, Kana-Kanji mixed string, corresponding to **b** and **c** of the correct data, respectively. Each of the following three cases is counted for the evaluation.

SS (Segmentation Success): **b'**= **b**
CS (Complete Success): **b'**= **b** and **c'**= **c**
TS (Tolerative Success): **b'**= **b** and **c'**~ **c**

There are many kinds of notational fluctuation in Japanese. For example, the conjugational suffix of some kind of Japanese verb is not always necessitated, therefore,売り上げ,売上げ and 売上 are all acceptable results for input うりあげ *uriage* (sales). Besides, a single word has sometimes more than one Kanji notations, e.g. 浜 *hama* (beach) and 濱 *hama* (beach) are both acceptable, and so on. **c'**~ **c** in the case of TS means that **c'** coincides with **c** completely or excepting the part which is heteromorphic in the above sense. For this, each of our conversion system has a dictionary which contains approximately 35,000 fluctuated notations of conceptual words.

4.3 Results of Experiments

Results of the experiments are given in Table 1 and Table 2 for input file f1 and f2, respectively.
Comparing the statistics of system A with D, we can conclude that the introduction of approximately 135,000 collocation data causes 8.1 % and 10.5 % raise of CS and TS rate, respectively, in case of relatively short input strings (f1). The raise of SS rate for f1 is 2.7%. In case of the longer input strings (f2) whose average number of bunsetsus is approximately 12.6, the raise of CS, TS and SS rate is 2.4 %, 5.2 % and 5.7 %, respectively. As a consequence, the raise of CS, TS and SS rate is 6.2 %, 9.1 % and 3.8 % on the average, respectively.

	System A	System B	System C	System D
SS(Segmentation Success)	9,656(90.0%)	9,912(92.4%)	9,927(92.5%)	9,954(92.7%)
CS(Complete Success)	5,085(47.4%)	5,638(52.5%)	5,677(52.9%)	5,953(55.5%)
TS(Tolerative Success)	6,226(58.0%)	6,971(64.9%)	7,024(65.4%)	7,355(68.5%)

Table 1:Result of the experiments for 10,733 short input strings data , f1.
(average number of Kana characters per input is 13.7)

	System A	System B	System C	System D
SS	8,345(68.4%)	8,978(73.6%)	8,988(73.7%)	9,037(74.1%)
CS	2,422(19.9%)	2,660(21.8%)	2,673(21.9%)	2,717(22.3%)
TS	3,965(32.5%)	4,555(37.4%)	4,568(37.5%)	4,601(37.7%)

Table 2: Result of the experiments for 12,192 long input strings data , f2.
(average number of Kana characters per input is 42.7)

	System D'	WXG
SS	9,949(92.7%)	9,804(91.3%)
CS	6,180(57.6%)	5,877(54.8%)
TS	7,646(71.2%)	7,290(67.9%)

Table 3:Comparison of system D' with WXG for f1.

	System D'	WXG
SS	8,928(73.2%)	8,815(72.3%)
CS	2,738(22.5%)	2,694(22.1%)
TS	4,649(38.1%)	4,543(37.3%)

Table 4: Comparison of system D' with WXG for f2.

4.4 Comparison with a Software on the Market

We compared System D with a Kana-to-Kanji conversion software for PC on the market, WXG Ver2.05 under the same condition except for the amount of installed collocation data. For this, system D was reinforced and renamed D', by equipping with WXG's 10,000 items of word dependency description. Both systems were disabled for the learning function. WXG has approximately 60,000 collocations (3,000 uninterruptible and 57,000 interruptible collocations), whereas System D' has approximately 135,000 collocations. The statistical results are given in Table 3 and Table 4 for the corpus f1 and f2, respectively.

The tables show that the raise of CS, TS and SS rate, which was obtained by System D' is 2.5 %, 4.5 % and 3.9 % on the average, respectively. No further comparison with the commercial products has been done, since we judge the performance of WXG Ver.2.05 to be average among them.

4.5 Discussions

Table 1 ~ 4 show that the longer input the system is given, the more difficult for the system to make the correct solution and the difference between accuracy rate of WXG and system D' is less for f2 than for f1. Further investigation clarified that the error of System D is mainly caused by missing words or expressions in the machine dictionary. Specifically, it was clarified that the dictionary does not have the sufficient number of Kata-Kana words and people's names. In addition, the number of fluctuational variants installed in the dictionary mentioned in **4.2** turned out to be insufficient. These problems should be remedied in future.

5. Concluding Remarks

In this paper, the effectiveness of the large scale collocation data for the improvement of the conversion accuracy of Kana-to-Kanji conversion process used in Japanese word processors was clarified, by relatively large scale experiments.

The extensive collection of the collocations has been carried out manually these ten years by the authors in order to realize not only high precision word processor but also more general Japanese language processing in future. A lot of resources, school textbooks, newspapers, novels, journals, dictionaries, etc. have been investigated by workers for the collection. The candidates for the collocation have been judged one after another by them.

Among collocations described in this paper, the idiomatic expressions are quite burdensome in the development of NLP, since they do not follow the principle of compositionality of the meaning. Generally speaking, the more extensive collocational data it deals with, the less the "rule system" of the rule based NLP system is burdened. This means the great importance of the enrichment of collocational data. Whereas it is inevitable that the arbitrariness lies in the human judgment and selection of collocations, we believe that our collocation data is far more refined than the automatically extracted one from corpora which has been recently reported [Church, K. W. et al, 1990, Ikehara, S. et al, 1996, etc.].

We believe that the approach described here is important for the evolution of NLP product in general as well.

References

Shudo, K. et al., 1980. Morphological Aspect of Japanese Language Processing. in Proc. of 8 th Internat.Conf. on Computational Linguistics(COLING80)

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. ,1986. Realization of Kana-to-Kanji Conversion Using Neural Networks. in Toshiba Review, 47-11. (in Japanese)

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

Shudo, K. et al. ,1988. On the Idiomatic Expressions in Japanese Language. in IPSJ SIG NL-66. (in Japanese)

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

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

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