Language, Information and Computation (PACLIC12), 18-20 Feb, 1998, 295-302

Japanese Kana-to-Kanji Conversion using Large Scale Collocation Data

Yasuo Koyama, Masako Yasutake, Kenji Yoshimura and Kosho Shudo
Fukuoka University
Fukuoka, 814-80 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

Japanese word processor or the computer used in Japan employs 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 homophones. In this paper, we report the results of our Kana-to-Kanji conversion experiments which embody the homophone processing using extensive collocation data. It is shown that approximately 135,000 collocation data yields 9.1 % raise of the conversion accuracy compared with the prototype system which has no collocation data.

1. Introduction

Japanese word processor or the computer used in Japan ordinarily employ the Japanese input method through keyboard stroke combined with Kana (phonetic) to Kanji (ideographic, Chinese) character conversion technology, because no extra technology such as the free hand character recognition or the speech recognition is required. 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 decide the substring of the input 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 most proper Kanji among many homophonic candidates.

The conventional methodology for processing the homophone has used the function to give the first priority to the word which was used lastly or to the word which is used most frequently. In fact, this method is effective in some situations, but sometimes tends to output the inadequate conversion result due to the lack of consideration on the semantic consistency of the concurrence of words. While it is difficult to employ the syntactic or semantic processing in earnest for the word processor from the cost vs. performance viewpoints, 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], Examining the consistency of the concurrence of adjacent words [Honma, S. 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 in case the keyword appropriate to the topic is detected in the input [Yamamoto, K. et al., 1992], Employing the validity of the concurrence of a noun and a verb which is calculated statistically [Takahashi, M. et al., 1996]. In any of these studies, where the main concern is to examine the consistency of word concurrence in the input, there are many problems left unsolved.

Besides these semantic or quasi-semantic gadgets, there seems to be the room for the improvement by using word level resources, namely, by the extensive use of the collocation, recurrent combination of words which co-occur more often than expected by chance. How many collocations should be collected and how much they contribute to the accuracy of Kana-to-Kanji conversion have not been reported yet.

In this paper, we present some results of our experiments of Kana-to-Kanji conversion, focusing on the usage of the 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

Contrary to the recent works on the automatic extraction of collocations from large corpus [Honma, S. et al, 1986, 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, putting years in it. This is because no stochastic framework assures the accuracy of the extraction, namely the necessity and sufficiency of the data set. The sparseness problem of the expression seems quite crucial for the stochastic approach. In addition, the validity check of the individual expression by human is inevitable in any way.

The collocations which we collected for our Kana-to-Kanji conversion system consist of two kinds: (1) idiomatic expressions (in a broad sense), 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 a broad sense) 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 (or dependent) words such as particles (postpositionals) or auxiliary verbs, the other is a class of **conceptual** collocations which work as independent words such 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 occasionally allow insertion of words between component words. In the following, the parenthesized number is the number of expressions adopted by the system.

2.1 Functional Collocations (2,174)

We call the expression which works like a particle **relational** collocation and the expression which works 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)
	に/おける	*ni/okeru* (at)
	た/ところで	*ta/tokorode* (even though)
	ものだから	*monodakara* (because)

auxiliary predicative collocations (1,414)

ex..	ければ/ならない	*nakereba/naranai* (must)
	ずに/いられない	*zuni/irarenai* (be obliged to)
	かも/しれない	*kamo/sirenai* (might)

2.2 Uninterruptible Conceptual Collocations (54,290)

four-Kanji-compound type (2,231)

ex.	我田引水	*gadeninsui* (every miller draws water to his own mill)
	千載一遇	*senzaiichiguu* (golden (opportunity))

adverb + particle type (3,089)

ex. あたふたと *atafutato* (disconcertedly)

近々に *kinkinni* (in the near future)

adverb + suru type (1,043)

ex. あくせくする *akusekusuru* (toil and moil)

ゆっくりする *yukkurisuru* (relax)

noun type (21,128)

ex. 赤の/他人 *akano/tanin* (perfect stranger)

袋の/鼠 *fukurono/nezumi* (rat in a trap)

verb type (13,225)

ex. おつりが/来る *otsuriga/kuru* (be enough to make the change)

一息/入れる *hitoiki/ireru* (take a break)

adjective type (2,394)

ex. 裏悲しい *uraganashii* (mournful)

油っこい *aburakkoi* (fatty)

adjective verb type (397)

ex. ご機嫌/斜め *gokigen/naname* (in a bad mood)

お誂え/向き *oatsurae/muki* (fit perfectly)

adverb and other type (8,185)

ex. 目に/見えて *meni/miete* (remarkably)

乗るか/反るか *noruka/soruka* (at all risks)

proverb type (2,598)

ex. 老いては/子に/従え *oiteha/koni/shitagae* (when old, obey your children)

馬の/耳に/念仏 *umano/mimini/nenbutsu* (a nod is as good as a wink to a blind horse)

2.3 Interruptible Conceptual Collocations (78,251)

noun type (7,627)

ex. 暖簾に/腕押し *orenni/udeosi* (to catch the wind with a net)

臭い物に/蓋 *kusaimononi/futa* (to hush up a scandal)

verb type (64,087)

ex. 後ろ髪を/引かれる *ushirogamiwo/hikareru* (feel as if one's heart were left behind)

足を/洗う *ashiwo/arau* (wash one's hands of a thing)

adjective type (3,617)

ex. 態度が/大きい *taidoga/ookii* (act in a lordly manner)

脇が/甘い *wakiga/amai* (be off guard)

adjective verb type (2,018)

ex. 役者が/上 *yakushaga/ue* (be more able)

形勢が/不利 *keiseiga/furi* (the situation is against)

others (902)

ex. 後に/引けぬ *atoni/hikenu* (can not give up)
間尺に/合わぬ *mashakuni/awanu* (unreasonable)

These collocations are not treated as single units but stored in the dependency file of the systems which prescribe semantically the validity of the concurrence of words.

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 the theoretical model, **bunsetsu** and **extended bunsetsu (e-bunsetsu)** [Shudo, K. et al., 1980] adopted as the unit of the segmentation of the input Kana string, and on its heuristic method, **minimum cost method** [Yoshimura,K. et al., 1987] based on a breadth first search algorithm for the reduction of the ambiguity of the segmentation.

A bunsetsu is well known basic postpositional or predicative phrase which composes Japanese sentence, defined as follows:
<bunsetsu>::=<conceptual word | numeral | formal noun |auxiliary declinable word | prefix | suffix | pre-numeral | post-numeral>
<functional word>*
<functional word>::=<particle | auxiliary verb>

An e-bunsetsu is defined as follows:
<e-bunsetsu>::=<prefix>* <conceptual word |uninterruptible conceptual collocation> <suffix>*
<functional word | functional collocation>*

Generally, e-bunsetsu is longer than bunsetsu, since it may include the collocation. More refined rules of the connectability of words to compose the bunsetsu and e-bunsetsu are used in the segmentation process. The interruptible conceptual collocation is not treated as a single unit but a string of bunsetsus by 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, and then segments it into 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 bunsetsus or e-bunsetsus in the example in this paper is denoted by "/".

ex. 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)
bunsetsu-segmentation:
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.
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 functional collocation and uninterruptible conceptual collocation tend to have priority over the ordinary word, and the e-bunsetsu does over the bunsetsu. 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 the specific e-bunsetsu boundary 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 causes the limitation of 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 the following conventional lexical resources:

functional words (1,010)

- particles (168)
- auxiliary verbs (54)
- formal nouns (23)
- auxiliary declinable words (47)
- prefixes (20)
- suffixes (236)
- pre-numeral (20)
- post-numeral (442)

conceptual words (131,661)

- nouns (97,737)
- verbs(14,690)
- adjectives (2,484)
- adjective verbs (9,166)
- adverbs (6,361)
- adnominals (326)
- interjections (485)
- connectives (289)
- others (123)

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 newly 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 newly with four-Kanji-compound type (2,231) and proverb type collocations(2,598).

System D is System C equipped newly 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, 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.

An example is shown below.

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 that is occasionally required by the difference of definition of "bunsetsu". 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, and then it is checked that the first (least cost) output Kana-Kanji mixed string of the system coincides with **b** and **c**. The system outputs two forms of the 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 the 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. Kata-Kana character , frequently used for the imported word in Japan, is treated as Kanji in the system. An imported word has sometimes more than one Kata-Kana transcriptions, depending on the fluctuation of its pronunciation, e.g. ヴァイオリン (violin) and バイオリン (violin) are both acceptable. **c'~ c** in the case of TS means that **c'** coincides with **c** except for 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 corpus 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 string (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.

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 is slightly revised, namely, D is newly equipped with10,883 items, prescribing the dependency relationship between conceptual words, which WXG is equipped with ,to be used in setting the priority of the

Table 1:Result of the experiments for 10,733 short input strings data , f1.
(average number of bunsetsus per input is 3.5)

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

Table 2: Result of the experiments for 12,192 long input strings data , f2.
(average number of bunsetsus per input is 12.7)

	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%)

segmentation and Kanji through he cost calculation. Both systems have no learning function. WXG has approximately 69,000 collocations (3,000 uninterruptible and 66,000 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 revised System D is renamed System D'. 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. We judge the performance of WXG Ver.2.05 to be average among them.

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

	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 4:Comparison of system D' with WXG in case of f2.

	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%)

4.5 Discussions

The effectiveness of a large scale collocation data for the improvement of the conversion accuracy of Japanese Kana-to-Kanji conversion system is confirmed. In addition to the raise of the accuracy, System D realized friendly user interface in the sense that the user, if necessary, can select Kanjis contained in a uninterruptible collocation, separately. This is because System D first regards the uninterruptible collocation as a single unit to give the priority to the (combination of) Kanjis in it, but can treat it as a string of bunsetsus, if the user expects. For example, input はらをたてる *harawotateru* is first regarded as a collocation 腹を/立てる (get angry) by the System D. However, 原を/立てる (let Mr. Hara go (to the batter's box)), or 原を/たてる is easily obtained, if necessary, because はらを *harawo* and たてる *tateru* are treated separately as well in the system.

The error of System D is mainly caused by the "unknown word", missing word or expression in the machine dictionary. Specifically, it was clarified by the experiments that the dictionary does not have 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 processor was clarified, by relatively large scale experiments.

The extensive collection of the collocations has been carried out manually for 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 of collocations. 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 the natural language processing, 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 natural language processing system is burdened. This means that it is quite important to enrich the collocational data further.

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 a large corpus which has been recently reported in some papers [Honma, S. et al, 1986, Church, K. W. et al, 1990, Ikehara, S. et al, 1996, etc.].

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)

Honma, S. et al. ,1986. Translation of Non-segmented Kana-Sentences to Kanji-Kana Sentences Using Collocation Information. in Trans. of IPSJ, 27-11. (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)

Curch, 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)

Takahashi, M. et al. ,1996. Processing Homonyms in the Kana-to-Kanji Conversion. in Proc. of 16th Internat. Conf. on Computational Linguistics (COLING 96)

Takahashi, M. et al. ,1996. Processing Japanese Homonyms Using Information about the Word Co-occurrence in the Simple Sentence. in Trans. of IPSJ, 36-6. (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)