# An HMM-based Method for Segmenting Japanese Terms and Keywords based on Domain-Specific Bilingual Corpora

**Keita Tsuji**
Library and Information Science Course, Graduate School of Education
The University of Tokyo, 7–3–1 Hongo, Bunkyo-ku, Tokyo, 113 Japan
E-Mail: i34188@m-unix.cc.u-tokyo.ac.jp

**and Kyo Kageura**
NACSIS, 3–29–1 Otsuka, Bunkyo-ku, Tokyo, 112 Japan
E-Mail: kyo@rd.nacsis.ac.jp

## 1 Introduction

In this paper we introduce a simple HMM-based method for segmenting Japanese complex terms and keywords, using domain-dependent Japanese-English bilingual corpora. Our overall aim is to automatically extract translation rules for terms and use them for IR query translation; the method described here is a preliminary element of our longer-term goal.

Much effort has been devoted to Japanese morphological analysis (Matsumoto et al 1996; Nagata 1994; Papageorgiou 1994; Yamamoto & Masuyama 1997). The general performance levels typically achieved are quite good, but for the purpose of processing complex terms and keywords they have some problems and limitations. Firstly, most general purpose morphological analysers do not segment complex words consistently[1]. There are a few morphological analysers (e.g. Takeda & Fujisaki 1987) which are specialised for the analysis of units of Chinese-origin, but they cannot be easily extended for other units such as Katakana. In addition, in the IR-oriented applications which we have in mind, performance levels of the systems when used with pre-defined dictionaries tend to be much lower than their average performance, because of the high degree of domain-specificity. A more effective method of segmenting complex Japanese lexical units is needed as a precondition for extracting term translation patterns.

Based on these considerations, we adopt the following strategies in designing the method of segmentation: (1) use of a domain-dependent list of Japanese-English term/keyword pairs as the target of analysis, in which the English counterparts are used to decide the cardinality of the desired segmentation; (2) to apply statistical training to each input corpus and analyse the results, i.e. on-spot training and closed data analysis, assuming that a certain amount of domain-dependent bilingual data is available at once[2].

## 2 Method of Segmentation

### 2.1 Basic Definition of HMMs

The basic idea of segmenting Japanese terms and keywords is very simple. Assuming that the English/Japanese term or keyword pairs have correspondences at the level of their constituent elements, the Japanese part is segmented into the same number of units as English counterparts. For instance, '情報検索 (information retrieval)' will be segmented into two.

In order to obtain the correct segmentation under this assumption, a simple HMM-based method is used, in which the states are defined on the basis of characters[3], and all possible segmentation patterns are tried and the best path is found out on the basis of the standard HMM search algorithms.

The basic configuration of our HMM is shown in Figure 1, where K-n represents the label assigned to the n-th character of a unit, and IF represents

[1] In JUMAN 3.2, for instance, a consecutive Katakana sequence is simply regarded as a single unit.

[2] In this respect, our starting-point is similar to those of Kageura (1997) and Moriwaki et. al. (1996).

[3] The character-based HMM is commonly adopted for Japanese morphological analysis, e.g. Papageorgiou (1994), Takeda & Fujisaki (1987), Yamamoto & Maruyama (1997).

a label of initial/final state of a unit[4].

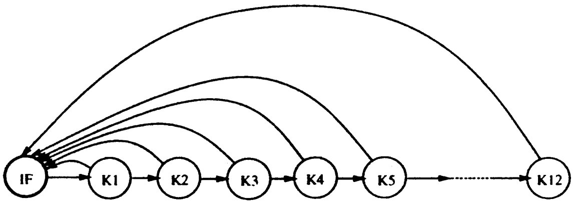

Fig. 1: Basic States/Transitions of the HMM for Complex Term Segmentation

For a 3-character word, for instance, there are three candidate sequences:

(1) <IF, K1, K2, K3, IF>
(2) <IF, K1, IF, K1, K2, IF>
(3) <IF, K1, K2, IF, K1, IF>

For example, (2) represents a 3-character word consisting of 2 units, i.e. a 1-character unit and a 2-character unit in this order.

As our aim is to extract Japanese units which correspond to English words, only those sequences which segment Japanese into the same number of English words are assigned to the data. So, in the case of Japanese-English pair '木構造' (tree structure)', only the candidates which decompose '木構造' into two units are considered, i.e. (2) and (3) above. The actual state sequences which are assigned to the label sequences (2) and (3), in this case, are as follows ($\phi$ represents a null character):

(2') <IF/$\phi$,K1/木,IF/$\phi$,K1/構,K2/造,IF/$\phi$>
(3') <IF/$\phi$,K1/木,K2/構,IF/$\phi$,K1/造,IF/$\phi$>

From these, state transitions such as (IF/$\phi$, K1/木), (K1/木, IF/$\phi$), (IF/$\phi$, K1/構), etc. are obtained, to which the probabilities are assigned based on the target corpus, using the forward-backward algorithm (Rabiner 1989; Charniak 1993).

The task of segmentation is to find out the state sequence with the maximum probability for each character sequence. Formally:

$$\arg\max_{S1,S2...Sn} p(S1, S2...Sn|C1, C2...Cn)$$

for a character sequence 'C1...Cn', where 'S1...Sn' is the state sequence consisting of K1...Ki.

### 2.2 Consideration of Character Types

There are five different types of characters in Japanese, viz.: Kanji (characters originating from Chinese, mainly used for nouns, stems of verbs, adjectives, and adverbs), Katakana (mainly used for nouns originating from Western-languages), Hiragana (often used for postpositions, inflexions and suffixes), alphabets (frequently used in technical papers), and other characters (Arabic numerals, punctuation, etc).

In Japanese, most of the different character types do not co-occur in a linguistically valid unit. Thus we should be able to gain a performance increment by incorporating segmentation preferences based on character types. The segmentation preferences are incorporated as an external heuristic operating on the probability assignment of the HMM; we multiply 1/10000000 by the probability of the state sequence which has transition (K-n/C1, K-(n+1)/C2) where C1 and C2 represent different character types[5], in order to make its probability relatively low.

## 3 Experiments and Results

### 3.1 Experiments

To evaluate the performance of the method, four data sources were used: (1) a list of artificial intelligence terms, extracted from Shapiro & Eckroth (1991) (henceforth AIT); (2) a list of documentation terms, extracted from Wersig & Neveling (1984) (DCT); (3) a list of keywords in artificial intelligence, extracted from the Database of Academic Conference Papers provided by NACSIS (AIK); and (4) a list of keywords in forestry, extracted from the same database (FRK).

For each of the four sources, the method described in the previous section was applied, and the results manually evaluated. To provide a baseline for comparison, we also applied a general purpose morphological analyser, JUMAN version 3.2, to the same data and evaluated the results.

### 3.2 Results

Table 1 lists the results of the experiments. In Table 1, 'Corr.pairs' indicates the number of pairs whose Japanese constituents and English words completely correspond to each other. 'CSP' in-

[4] We introduce restrictions that the segmented units should not be longer than 12 characters, and that one-character units should not appear more than twice in succession.

[5] However, when C2 is Hiragana, we do not apply this procedure, because Hiragana characters are often used for representing inflection.

dicates the number of correctly segmented pairs[6]. 'Ratio A' and 'Ratio C' indicate the percentages of the correctly segmented pairs against all pairs and corresponding pairs respectively. The row '2+' shows the pairs in which English entries consist of more than two words. The column 'HMM (basic)' and 'HMM with CC' show the results of our methods without and with character type constraints, respectively.

In the case of JUMAN, the results are regarded as correct when there are JUMAN segmentations corresponding to English words and at the same time the other segmentations reflect proper Japanese components. So for instance, '因果/的/順序/づけ (causal ordering)' is evaluated as correct because the segmentation between '的' and '順序' corresponds to the English counterparts while the other segmentations, i.e. between '因果' and '的' on the one hand and '順序' and 'づけ' on the other properly reflect Japanese components[7].

The evaluation of the results by JUMAN are done only for the pairs whose English entries consist of more than two words, and whose Japanese constituents and English words correspond completely, because they are our main concern, and are considered to reflect the general performance of JUMAN, as JUMAN does not refer to English counterparts anyway.

### *3.3 Observations*

Firstly, compared to the performance of JUMAN, we observed that the HMM method gives better results among the corresponding pairs, although the performance depends to some extent on the data. We also observed that the HMM method with character type considerations gives consistently better performances than the HMM method which does not incorporate preferences for character types.

Secondly, we can observe that the overall performances differ considerably when run on different types of data. On the lists of terms, AIT and DCT, the HMM method with character type considerations correctly segments more than 87% of all the pairs with more than two English words. On the other hand, reflecting the low ratio of corresponding pairs in the data, the performance is considerably lower for the lists of keywords (AIK and FIK). At this level, therefore, the nature of the data greatly influences performance.

If we evaluate the performance against the number of corresponding pairs, we can observe that the results are quite satisfactory, i.e. well over 90% correct segmentation among the pairs with more than two-units in AIT, DCT and AIK. Although the performance in FRK is still low, the difference of the performance with other data is much smaller.

## 4 Conclusion

We conclude that our method gives relatively high performance, as long as the Japanese and English pairs have correspondence at the level of their constituent elements. In this respect, on-spot training with a small training data for complex HMM state-transitions performs fairly well. Although the assumption of closed training/processing limits the range of applications, it is fairly useful, especially when coping with highly specialised or domain-dependent data. On the other hand, we can observe that the result is highly dependent on the domain and/or the type of data (term list or keyword list).

In service of our eventual goal as described in the first section, we plan to enhance the method in the following ways:

(1) extracting the English-Japanese pairs at the same time, by incorporating the possible combinations of English words into the state transition patterns;

(2) distinguishing those pairs whose English and Japanese correspond at the level of constituent units from those whose constituent units do not correspond;

(3) decomposing English words into morphemes, which correspond to Japanese units, such as M' unit, and extracting morpheme level pairs;

(4) calculating the transition probabilities on the basis of correctly segmented units.

In fact, we have already extended the method to incorporate (1) and (2) and have obtained fairly good results. Also, we expect that the basic performance can be further improved by fine-tuning

[6] We regard the results as correct only when all the Japanese segmented elements correspond to English counterparts.

[7] It should be emphasised that it is not fair to evaluate JUMAN from this point of view, because segmenting complex terms and keywords into units which correspond to English are not primary goals of JUMAN. So the figures by JUMAN are given here only for reference.

| Data | Engl. | All | Corr. | HMM (basic) | | | HMM with CC | | | JUMAN | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Word | Pairs | Pairs | CSP | Ratio A | Ratio C | CSP | Ratio A | Ratio C | CSP | Ratio C |
| AIT | 1 | 938 | 938 | 938 | 100.00 | 100.00 | 938 | 100.00 | 100.00 | | |
| | 2 | 2222 | 2167 | 1965 | 88.43 | 90.68 | 2036 | 91.63 | 93.95 | 1887 | 87.08 |
| | 3 | 475 | 428 | 349 | 73.47 | 81.54 | 370 | 77.89 | 86.45 | 368 | 85.98 |
| | 4+ | 109 | 80 | 65 | 59.63 | 81.25 | 66 | 60.55 | 82.50 | 73 | 91.25 |
| | Total | 3744 | 3613 | 3317 | 88.60 | 91.81 | 3410 | 91.08 | 94.38 | | |
| | 2+ | 2806 | 2675 | 2379 | 84.78 | 88.93 | 2472 | 88.10 | 92.41 | 2328 | 87.03 |
| DOC | 1 | 393 | 393 | 393 | 100.00 | 100.00 | 393 | 100.00 | 100.00 | | |
| | 2 | 718 | 671 | 633 | 88.16 | 94.34 | 642 | 89.42 | 95.68 | 585 | 87.18 |
| | 3 | 64 | 52 | 48 | 75.00 | 92.31 | 47 | 73.44 | 90.38 | 44 | 84.62 |
| | 4+ | 13 | 4 | 4 | 30.77 | 100.00 | 4 | 30.77 | 100.00 | 2 | 50.00 |
| | Total | 1188 | 1120 | 1078 | 90.74 | 96.25 | 1086 | 91.41 | 96.96 | | |
| | 2+ | 795 | 727 | 685 | 86.16 | 94.22 | 693 | 87.17 | 95.32 | 631 | 86.80 |
| AIK | 1 | 656 | 656 | 656 | 100.00 | 100.00 | 656 | 100.00 | 100.00 | | |
| | 2 | 1726 | 1606 | 1482 | 85.86 | 92.28 | 1531 | 88.70 | 95.33 | 1379 | 85.87 |
| | 3 | 485 | 382 | 309 | 63.71 | 80.89 | 340 | 70.10 | 89.00 | 301 | 78.80 |
| | 4+ | 124 | 53 | 40 | 32.26 | 75.47 | 46 | 37.10 | 86.79 | 45 | 84.91 |
| | Total | 2991 | 2697 | 2487 | 83.15 | 92.21 | 2573 | 86.02 | 95.40 | | |
| | 2+ | 2335 | 2041 | 1831 | 78.42 | 89.71 | 1917 | 82.10 | 93.92 | 1725 | 84.52 |
| FRK | 1 | 692 | 692 | 692 | 100.00 | 100.00 | 692 | 100.00 | 100.00 | | |
| | 2 | 1685 | 1242 | 989 | 58.69 | 79.63 | 1081 | 64.15 | 87.04 | 990 | 79.71 |
| | 3 | 502 | 279 | 203 | 40.44 | 72.76 | 211 | 42.03 | 75.63 | 171 | 61.29 |
| | 4+ | 159 | 29 | 18 | 11.32 | 62.07 | 20 | 12.58 | 68.97 | 21 | 72.41 |
| | Total | 3038 | 2242 | 1902 | 62.61 | 84.83 | 2004 | 65.96 | 89.38 | | |
| | 2+ | 2346 | 1550 | 1210 | 51.58 | 78.06 | 1312 | 55.92 | 84.65 | 1182 | 76.26 |

Table 1. Results of Experiments

the character-type combination preference heuristics.

## References

Charniak, E. (1993) *Statistical Language Learning.* Cambridge, Mass: MIT Press.

Kageura, K. (1997) "Moji Tan'i no Bigram Syakudo ni Motozuku Fukugou Kanjiretu no Tan'igiri Syuhou," *Proc. of the Third Annual Meeting of the Association for Natural Language processing.* p. 477–481. (in Japanese)

Matsumoto, Y., Kurohashi, S., Yamaji, O, Taeki, Y, and Nagao, M. (1997) *Japanese Morphological Analysis System JUMAN Manual.* ver. 3.2. (in Japanese)

Moriwaki, S., Kawabe, K. and Tsujii, J. (1996) "Jisyo wo Tukawanai Nihongo Senmon Yogo no Jidou Bunkatu," *Proc. of the Second Annual Meeting of the Association for Natural Language processing.* (in Japanese)

Nagata, M. (1994) "A Stochastic Japanese Morphological Analyzer using a Forward-DP Backward-A* N-best Search Algorithm," *COLING-94.* p. 201–207.

Papageorgiou, C. P. (1994) "Japanese Word Segmentation by Hidden Markov Model," *Proc. of Human Language Technology Workshop.* p. 283–288.

Rabiner, L. R. (1989) "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proceedings of IEEE,* 77(2) 257–295.

Shapiro, S. C. and Eckroth, D. (1991) *Jinko Tinou Daijiten.* Tokyo: Maruzen. [*Encyclopedia of Artificial Intelligence.* New York: Wiley. 1987.]

Takeda, K. and Fujisaki, T. (1987) "Automatic Decomposition of Kanji Compound Words Using Stochastic Estimation," *Transactions of Information Processing Society of Japan,* 28(9) 952–961. (in Japanese)

Yamamoto, M. and Masuyama, M. (1997) "Hinsi-Kugiri Joho wo Fukumu Kakucyomoji no Rensa Kakuritsu wo Motiita Nihongo Keitaiso Kaiseki," *Proc. of the Third Annual Meeting of the Association for Natural Language processing.* p. 421–424. (in Japanese)

Wersig, G. and Neveling, U. (eds.) (1984) *UNESCO Joho Kanri Yogosyu.* Tokyo: Maruzen. [*Terminology of Documentation.* Paris: Unesco. 1976.]