Extracting Morpheme Pairs from Bilingual Terminological Corpora

Keita TSUJI
Library and Information Science Course, Graduate School of Education,
The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-0033 Japan

Kyo KAGEURA
National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430 Japan

Abstract

An HMM-based method for extracting bilingual morpheme pairs from domain-specific bilingual term lists is reported in this paper. In our method, Japanese terms are segmented into morphemes while at the same time the corresponding English and Japanese morphemes are identified, by taking advantage of the nature of the bilingual terms. The method gives satisfying results for the two terminological data that we used for the evaluation.

1 Introduction

In this paper we introduce an HMM-based method for extracting English and Japanese morpheme pairs from domain-specific bilingual term lists. In our method, the task of segmenting Japanese terms is carried out while at the same time the corresponding English and Japanese morphemes are identified, on the basis of adaptive on-spot training. As the method requires no pre-processing, it has an advantage in the treatment of domain-specific terms.

There has recently been a growing interest in terminology related applications, and many term lists have become available in electronic form. In this context, the automatic identification of the corresponding morphemes in a term list is very useful, because it can further be used for many terminological applications. For example, once we obtain a list of morpheme pairs, it can be used as bootstrapping information for the automatic identification of translational term pairs from bilingual textual corpora. Or, it can be used for automatically extracting translation rules of complex terms. The work reported here constitutes an essential element of these wider applications.

2 Backgrounds

In order to extract the Japanese-English morpheme-word pairs (henceforth 'morpheme pairs' for simplicity) from term lists, two tasks are essential, i.e. (1) to segment the Japanese complex terms into proper constituent elements or morphemes, and (2) to identify the corresponding pairs.

With respect to the former, much effort has been devoted to Japanese morphological analysis including word segmentation, both for the general morphological analysis (Matsumoto et al., 1997; Nagata, 1994; Papageorgiou, 1994; Yamamoto & Masuyama, 1997) and for the analyses of specific morphological phenomena (Takeda & Fujisaki 1987).

Many of the proposed methods perform satisfactorily in the tasks that they intend to perform. However, in the task of segmenting complex terms, the performance is not totally satisfactory, for the following reasons:

(1) The main purpose of the general purpose morphological analysers is to recognise and tag the linguistic items for the purpose of analysing sentences, so most of them do not segment complex words consistently. On the other hand, there are a few analysers, e.g. Takeda and Fujisaki (1987), which decompose Kanji sequences into a consistent unit, but the method is not easily extensible to other character types nor to different unit sizes.

(2) As we treat domain-specific technical terminology, performance levels of the systems with pre-defined dictionaries tend to be lower than their average performance. It is possible to enhance the dictionary by domain-specific lexical elements. In that case, however, the task we are trying to perform may better be interpreted as a preliminary element of the dictionary enhancement, and not vice-versa.

It is preferable, therefore, not to assume any pre-processing, and to incorporate the task of segmentation into the procedure of identifying morpheme pairs. The domain-specificity also requires that our method should preferably be adaptive, combined with some general assumptions.

As for the identification of translation pairs, there have recently been a lot of studies. Methodologically, the studies can be categorised into two types. The first is based on some summary/test statistic measures such as the mutual information or ϕ^2 (Gale & Church, 1991; van der Eijk, 1993; Fung, 1995;

Smadja et al., 1996), and the other is based on the ideas of some kinds of optimisations, which methodologically use some recursive calculations (Kupiec, 1993; Melamed, 1996).

When extracting the translational word pairs from bilingual textual corpora, it is in general necessary first to align sentences and define corresponding units which are often bigger than one sentence. So the straightforward application of such well established methods as HMM is computationally prohibitive, even if a suitable model can be defined. In our case, however, the size of the individual term list is not expected to be very big, and term-level correspondences are assumed on the side of the data. The computational cost, therefore, is estimated not to be prohibitively high for some moderately costly methods. In our study, we have chosen the HMM-based method, as we recognise it to be conceptually suitable to the task of segmenting and identifying corresponding morpheme pairs.

3 The Method

3.1 Basic Assumptions

The basic idea of segmenting Japanese terms and identifying corresponding morpheme pairs is simple and straightforward. Firstly, we adopt the following two simple assumptions:

(1) A Japanese term consists of the same number of morphemes as the number of words which constitute the corresponding English term.

(2) For each of the morphemes of a Japanese term, a corresponding English word exists among the constituent words of the corresponding English term.

These assumptions hold for most term pairs. For instance, the term '>pJs8!:w(information retrieval)' consists of two morphemes, i.e. '>pJs' and '8!:w' which correspond to 'information' and 'retrieval', respectively.

Because the number of constituent elements in a term is limited, the task of identifying corresponding pairs is not very complex, if the morphemes are identified in advance. What makes the problem more complex here is the fact that the constituent morphemes of Japanese terms are not defined in advance. Thus the task here, in the above example, is to choose the correct pairs among the following six possibilities.

```
!>p (information) + Js8!:w (retrieval)
!>pJs (information) + 8!:w (retrieval)
!>pJs8! (information) + :w (retrieval)
!>p (retrieval) + Js8!:w (information)
!>pJs (retrieval) + 8!:w (information)
!>pJs8! (retrieval) + :w (information)
```

Assuming that, in a terminology of a domain, the same concept is basically represented by the same morpheme or word, it is expected to be possible to identify the proper translation pairs, using the information contained in a given corpus. Because the formation of complex terms is mostly systematic (Sager, 1990), we expect that the necessary information can be extracted from relatively small term lists, by adaptive training.

3.2 Definition of HMM

The task of segmenting Japanese morphemes and identifying corresponding morpheme pairs can be performed in various ways. Here we adopt an HMM-based method, because (1) the HMM has been proved to be particularly useful in the task of decomposing Japanese sequences (Papageorgiou, 1994; Takeda & Fujisaki, 1987), (2) the task of identifying corresponding English words can be straightforwardly incorporated into the model, and (3) it is computationally reasonable to analyse terminological data by the HMM.

In our HMM, we define the states on the basis of individual Japanese characters, not of words. In this respect, our definition of HMM is similar to those of Papageorgiou (1994), Takeda & Fujisaki (1987) and Yamamoto & Masuyama (1997). Assuming that the length of a Japanese morpheme does not exceed 12 characters, the basic states/transitions of the HMM can be defined as in Figure 1, where K-n represents the label assigned to the n-th character of a unit or a possible morpheme, and IF represents a label of initial/final state of a unit.

In order to identify the translation pairs, a unique English candidate word is assigned to each unit delimited by the IF state.

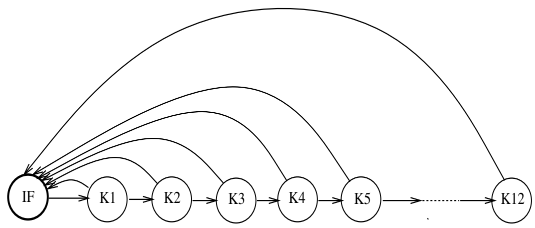

Figure 1: Basic Configurations of the HMM

In case of the Japanese-English term pair 'LZ9=B$ (tree structure)', for instance, the following four state sequences are created (ϕJ and ϕE represent a null character):

(1) (ϕJ/ϕE/IF) $\rightarrow$ (LZ/tree/K1) $\rightarrow$ (ϕJ/ϕE/IF) $\rightarrow$ (9=/structure/K1) $\rightarrow$ (B$/structure/K2) $\rightarrow$ (ϕJ/ϕE/IF)

(2) (ϕJ/ϕE/IF) → (LZ/structure/K1) → (ϕJ/ϕE/IF) → (9=/tree/K1) → (B$/tree/K2) → ($\phi$J/$\phi$E/IF)

(3) (ϕJ/ϕE/IF) → (LZ/tree/K1) → (9=/tree/K2) → (ϕJ/ϕE/IF) → (B$/structure/K1) → ($\phi$J/$\phi$E/IF)

(4) (ϕJ/ϕE/IF) → (LZ/structure/K1) → (9=/structure/K2) → (ϕJ/ϕE/IF) → (B$ /tree/K1) → (ϕJ/ϕE/IF)

Note that in all the sequences, the English parts are the same among the states that belong to the same unit, surrounded by the IF states.

From these, individual state transitions such as [(ϕJ/ϕEϕ/IF) → (LZ/tree/K1)], [(LZ/tree/K1) → (ϕJ/ϕE/IF)],..., [(B$/tree/K1) → ($\phi$J/$\phi$E/IF)] are obtained, to which the probabilities are assigned on the basis of data from the term list, using the forward-backward algorithm (Rabiner, 1989; Charniak, 1993).

The task of segmenting Japanese terms and identifying the corresponding morpheme pairs is carried out by finding the state sequence with the maximum probability for each Japanese character sequence or term, i.e.,

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

for a character sequence 'C1...Cn', where 'S1...Sn' is the state sequence, each state of which represents the triple (Japanese sequence/English word/state label). In the above example of 'LZ9=B$', we expect that the transition (1), which represents the correct segmentation and correspondence, has the maximum transition probability.

4 Basic Experiments

4.1 Data and Evaluation

Two terminological data are used in order to evaluate the performance of the method, i.e. a bilingual list of terms in artificial intelligence (Shapiro & Eckroth, 1991) and a bilingual list of terms in documentation (Wersig & Neveling, 1984).

Table 1 shows the basic quantities of the two lists of terms. The column 'No. of Words' indicates the total number of English words or constituent elements in the lists. 'Corresp. Terms' and 'Ratio of Corresp.' indicate the number of terms whose English and Japanese constituent elements completely correspond, and the ratio of these terms with respect to the total number of terms, respectively. The row 'All' indicates the number of terms including one-word terms, while '2+' indicates the number of complex terms. Note that, being a list of term types, the size is very small, compared to the standard textual corpora.

Data		No. of Terms	No. of Words	Corresp. Terms	Ratio of Corresp.
AI	All	3742	7112	3506	93.69
	2+	2804	6174	2568	91.58
Doc.	All	1188	3437	1116	93.94
	2+	792	2777	720	90.91

Table 1. Basic Quantities of the Data

Table 2 shows a small part of the list of artificial intelligence terms.

!35G0E*Lr3d	conceptual role
!35G0E*M=B,	conceptual expectation
!35G0$N%?%$%W	concept type
!35G0I=8=	concept representation

Table 2. Example of the Data

The HMM-based method described in the previous section was applied to each of the data. The results are evaluated for all the terms in the data, according to the following criteria:

(1) At the level of meaning, a pair is evaluated to be valid when (i) the morpheme pair exists as an independent term in the data, or (ii) the translation can be found in a standard English-Japanese dictionary (Koine, 1980).

(2) At the grammatical level, a pair is evaluated to be incorrect in the following cases:

- If the English word of the pair is a noun and the Japanese morpheme contains a suffix or a postposition, they are evaluated to be incorrect, e.g. '35G0$N (concept)' → × (cf. '35G0 (concept)' → ◯).
- If the English word is not a noun and the Japanese morpheme does not contain a suffix or a postposition, they are evaluated to be incorrect, e.g. '35G0 (conceptual)' → × (cf. '35G0$N (conceptual)' → ◯).

A pair is evaluated to be valid when it satisfies both the meaning and the grammatical criteria.

Table 3 shows the results of the two domains, by both the token-based and type-based evaluations. In the table, 'MP' indicates the number of the morpheme pairs contained in the lists, which equals the number of English words (see Table 1). 'EP' represents the number of morpheme pairs which are extracted by the HMM-based method, regardless of their correctness. 'EMP' represents the number of the correct morpheme pairs among the extracted pairs. 'Pre' and 'Rec' indicate the precision and the recall respectively, which are defined as follows:

$$Pre = 100 \times \frac{EMP}{EP}$$
$$Rec = 100 \times \frac{EMP}{MP}$$

As in Table 1, the row 'All' indicates the results calculated by all the terms in the lists including one word terms, while '2+' in the table represents the results of word pairs whose English entries consist of more than two words.

Artificial Intelligence						
		MP	EP	EMP	Pre	Rec
Token	All	7112	7246	6424	88.66	90.33
	2+	6174	6308	5486	86.97	88.86
Type	All	3437	3715	2984	80.32	86.82
	2+	2777	3047	2317	76.04	83.44
Documentation						
		MP	EP	EMP	Pre	Rec
Token	All	2012	2060	1869	90.73	92.89
	2+	1616	1664	1473	88.52	91.15
Type	All	1046	1128	966	85.64	92.35
	2+	758	839	676	80.57	89.18

Table 3. The Results of the Basic HMM Method

According to the token-based evaluation, the method achieves well over 85% precision and recall. The result is fairly satisfactory, considering the small size of the data as well as the fact that the accuracy of decomposing technical terms by a general purpose morphological analyser is around 85% to 90% on the basis of tokens (Tsuji & Kageura, 1997).

4.2 Comparative Experiment

For the purpose of further evaluation, we carried out a comparative experiment, applying the method which belongs to the first approach mentioned in Section 2. Instead of applying the mutual information (Fung, 1995) or ϕ^2 (Gale & Church, 1991), we used the likelihood ratio test, because it is properly sensitive to the low frequency data (Agresti, 1996; Dunning, 1993).

In order to apply the likelihood ratio test, the following contingency table is first defined:

		English Word		Total
		w_e+	w_e-	
Japanese	w_j+	f_{11}	f_{12}	$f_{1.}$
Sequence	w_j-	f_{21}	f_{22}	$f_{2.}$
	Total	$f_{.1}$	$f_{.2}$	$f_{..}$

where f_{11} indicates the frequency of a Japanese sequence w_j coupled with an English word w_e, f_{12} indicates the frequency of the Japansese sequence w_j coupled with any of the English words other than w_e, etc.

The likelihood ratio λ between w_j and w_e is defined as follows:

$$\begin{aligned} -2\log\lambda &= 2[\sum_c \log L(f_{1c}/f_{.c}, f_{1c}, f_{.c}) \\ &\quad - \sum_c \log L(f_{1.}/f_{..}, f_{1c}, f_{.c})] \end{aligned}$$

where

$$\log L(p,n,k) = k\log p + (n-k)\log(1-p)$$

For instance, the term for the first column of the table is calculated by:

$$\log L(\frac{f_{11}}{f_{.1}}, f_{11}, f_{.1}) = f_{11}\log\frac{f_{11}}{f_{.1}} + f_{21}\log\frac{f_{21}}{f_{.1}}$$

In order for the direction of bias of the table to be reflected, we multiplied the sign of Yule's Y (Agresti, 1996), defined for the same contingency table. Yule's Y is defined by the following equation:

$$Y = \frac{\sqrt{a}-1}{\sqrt{a}+1}$$

where

$$a = \frac{f_{11}\cdot f_{22}}{f_{12}\cdot f_{21}}$$

The identification of the corresponding pairs is carried out as follows. Firstly, extract all the possible Japanese-sequence/English-word combinations. For instance, in a term pair 'LZ9=B$ (tree structure)', the following four patterns are recognised, from which the eight possible pairs are extracted.

```
(1) <LZ(tree)>  <9=B$(structure)>
(2) <LZ(structure)>  <9=B$(tree)>
(3) <LZ9=(tree)>  <B$(structure)>
(4) <LZ9=(structure)>  <B$(tree)>
```

The frequency of each pair is counted, with the number of alternative elements in a term taken into consideration. So for instance, in the above case, the frequencies of 'LZ' and 'tree' are counted as 1/4 and 1/2 respectively, and the pair 'LZ (tree)' is counted as 1/8.

For each of the term pairs, then, the best segmentation/correspondence pattern is identified by choosing the largest value of the candidates, whose values are calculated by multiplying the scores of the constituent pairs.

Table 4 gives the results of the likelihood ratio based method. They show that the performance of the HMM-based method is significantly better. This clearly shows the effectiveness of the HMM-based method.

Artificial Intelligence						
		MP	EP	EMP	Pre	Rec
Token	All	7112	7246	5656	78.06	79.53
	2+	6174	6308	4718	74.79	76.42
Type	All	3437	3794	2465	64.97	71.72
	2+	2777	3137	1806	57.57	65.03
Documentation						
		MP	EP	EMP	Pre	Rec
Token	All	2012	2060	1729	83.93	85.93
	2+	1616	1664	1333	80.11	82.49
Type	All	1046	1136	854	75.18	81.64
	2+	758	847	564	66.59	74.41

Table 4. The Results of the Likelihood Based Method

5 Improvements of the Method

The examination of the pairs wrongly analysed by the HMM method revealed that there are basically two types of failure. One is the failure caused by the imbalance between the number of Japanese morphemes and English words. For instance, in some term pairs, the English terms may contain words which do not correspond to any of the morphemes of the Japanese terms in the first place. The other failure is caused by the incorrect segmentation of Japanese sequences. Many of the errors belonging to the former type can be avoided by ignoring the English stop words, while the errors of the latter type can be avoided to some extent by using some general heuristics in the recognition of Japanese sequences.

In order to improve the performance, therefore, we introduced a few general heuristics, which can be categorised into three types. Firstly, the following pre-processing procedure is applied to the English terms, before the HMM method is applied:

(1) Ignore 'a' and 'the'.

(2) Ignore 'of', unless the corresponding Japanese term contains '$N'.

(3) Ignore 's', 'or', 'ion', 'al' at the end of the word, e.g. 'constructor' is treated as 'construct'.

(4) Ignore 'ing' at the end of the word. If the same character appears twice consecutively before 'ing', ignore one of them. For instance, 'learning' is treated as 'learn', while 'mapping' is treated as 'map'.

(5) Replace the upper-case with the lower-case, unless the Japanese counterpart contains the same upper-case letters. For instance, 'Naive' in 'D>4QJ*M}3X@k8@ (Naive Physics Manifesto)' is changed to 'naive', while 'Mach' in 'Mach %P%s%I (Mach band)' is not changed.

Secondly, based on the considerations of valid Japanese sequences in general, we excluded the following types of pairs from the candidates.

(1) Morpheme pairs whose Japanese part begins with the following characters:

!$!!$#!$%!$'!$)!$c!$e!$g!$C!$$s!$
!%!!%#!%%!%'!%)!%c!%e!%g!%C!%%s!$!<

Or ends with the characters:

!$C!%C!$IT!$Hs

(2) Morpheme pairs whose Japanese part begins with a Hiragana or Katakana character, followed by characters of the other types, e.g. '$J8!:w'.

(3) Morpheme pairs whose Japanese part ends with one Katakana character, preceded by characters of the other types, e.g. 'O@M}%W'.

(4) Morpheme pairs whose Japanese part consists of only one Katakana character, e.g. '%i'.

These heuristics, of excluding improbable pairs and of giving preference to the possibly desirable pairs, are incorporated into the HMM method and change the probabilities of the relevant pairs.

Artificial Intelligence						
		MP	EP	EMP	Pre	Rec
Token	All	7112	7215	6646	92.11	93.45
	2+	6174	6277	5708	90.94	92.45
Type	All	3437	3622	3097	85.51	90.11
	2+	2777	2959	2435	82.29	87.68
Documentation						
		MP	EP	EMP	Pre	Rec
Token	All	2012	2055	1911	92.99	94.98
	2+	1616	1659	1515	91.32	93.75
Type	All	1046	1108	984	88.81	94.07
	2+	758	818	693	84.72	91.42

Table 5. The Results of the Improved HMM Method

Table 5 shows the results of the improved HMM method, applied to the same terminological data. Based on the token frequencies, both the precision and the recall are now well over 90%, even for '2+'. Based on the types, we achieved well over 80% precision and 85% recall. In case of documentation, the precision is almost 85% and the recall is over 90%.

The results show a great improvement over the basic HMM. As the heuristics introduced here are fairly general and completely independent of any particular domain, the performance improvement observed here in the two lists can be expected in other data of different domains.

6 Conclusions

We have introduced the HMM-based method of identifying Japanese/English morpheme pairs. The experiments show that it performs fairly well. The advantages of the method are:

(1) The segmentation of Japanese terms is carried out at the same time as the identification of the corresponding Japanese/English morpheme pairs.
(2) The method functions on the basis of adaptive on-spot training, assuming a certain amount of the terminological data, which may well be small.

A practical implication of the above is that the method is expected to perform equally well with any list of any subject domain, without any modifications.

In fact, there still is some room for further improvement, because the external heuristics introduced above may be improved further. In addition, we will be able to improve the performance further if we can formalise the method of distinguishing the terms whose constituent elements correspond, from those whose elements do not correspond, because, without the consideration of this, we cannot perform with 100% precision (see Table 1).

At the same time, we are currently exploiting possible applications, in the two directions described at the beginning, i.e. to extract translation rules of terms from terminological lists, and to apply the identified morpheme pairs to extract possible terminological pairs from bilingual textual corpora.

References

Agresti, A. (1996) *An Introduction to Categorical Data Analysis*. New York: John Wiley.

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

Dunning, T. (1993) "Accurate methods for the statistics of surprise and coincidence," *Computational Linguistics*. 19(1), p. 61–74.

Eijk, van der P. (1993) "Automating the acquisition of bilingual terminology," *Proc. of the 6th EACL*. p. 113–119.

Fung, P. (1995) "A pattern matching method for finding noun and proper noun translations from noisy parallel corpora," *Proc. of 33rd ACL*. p. 233–236.

Gale, W. A. and Church, K. W. (1991) "Identifying word correspondences in parallel texts," *Proc. of DARPA Speech and Natural Language Workshop*. p. 152–157.

Koine, Y. et al. (eds.) (1980) *Kenkyusha's New English-Japanese Dictionary*. 5th edition. Tokyo: Kenkyusha.

Kupiec, J. (1993) "An algorithm for finding noun phrase correspondences in bilingual corpora," *Proc. of 31st ACL*. p.17–22.

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

Melamed, I. D. (1996) "Automatic construction of clean broad-coverage translation lexicons," *2nd Conference of the Association for Machine Translation in the Americas*.

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 the IEEE*. 77(2), p. 257–286.

Sager, J. C. (1990) *A Practical Course in Terminology Processing*. Amsterdam: John Benjamins.

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

Smadja, F., McKeown, K. R. and Hatzivassiloglou, V. (1996) "Translating collocations for bilingual lexicons: A statistical approach," *Computational Linguistics*. 22(1), p. 1–38.

Takeda, K. and Fujisaki, T. (1987) "Automatic decomposition of Kanji compound words using stochastic estimation," *Transactions of Information Processing Society of Japan*, 28(9), p. 952–961. (in Japanese)

Tsuji, K. and Kageura, K. (1997) "An HMM-based method for segmenting Japanese terms and keywords based on domain-specific bilingual corpora," *NLPRS'97*. p. 557–560.

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

Yamamoto, M. and Masuyama, M. (1997) "Hinsikugiri 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)