Pacific Association for Computational Linguistics

CORRECTION OF WORD SEGMENTATION ERRORS THROUGH CHARACTER-BASED HMM

HIROYUKI SHINNOU AND MASANORI IKEYA
Ibaraki University, Department of Systems Engineering
4-12-1 Nakanarusawa, Hitachi, Ibaraki, 316-8511, Japan

Word segmentation for a compound is an important technology. In general, word segmentation can be conducted by morphological analysis. However, morphological analysis has the problem of unknown words. Instead character-based HMM does not suffer its problem. On the other hand, character-based HMM cannot recognize a word composed of many characters as one word, but morphological analysis can do it easily. Thus, the respective weaknesses of morphological analysis and character-based HMM compensate each other. In this paper, we propose a method in which morphological analysis and character-based HMM are complementary to each other, in order to segment a compound into words. We use character-based HMM to modify word segmentation errors produced by morphological analysis. To find these errors we use the error pattern of morphological analysis. Through experiments, we estimate our method can correct about three fourths of errors which morphological analysis will produce.

Key words: character-based HMM,morphological analysis,word segmentation

INTRODUCTION

In this paper, we propose a method in which general morphological analysis and character-based HMM (Hidden Markov Model) are complementary to each other, in order to segment a compound into words.

Word segmentation for a compound is an important technology because it is required not only in conventional natural language systems but also in indexing of full text search and retrieval and analysis of retrieval keys. In general, word segmentation can be conducted by morphological analysis. However, morphological analysis has the problem of unknown words besides incomplete accuracy. Thus, methods other than morphological analysis have been proposed for word segmentation, such as character-based HMM. Character-based HMM does not suffer the problem of unknown words because it is a character-based method. However, character-based HMM cannot recognize a word composed of many characters as one word if it learns costs to output a character with movement from one state to another by using uni-gram and bi-gram obtained from training data. Therefore, character-based HMM needs further development. Yamamoto (1997) added information of a part of speech to each character. Tsuji (1997) derived the number of words in a compound from a bilingual dictionary. Oda (1998) extended output symbols on HMM to variable length n-gram through PPM^* model.

we should note that morphological analysis does not suffer the above problem of character-based HMM. It is rather easy for morphological analysis to recognize a word composed of many characters as one word. On the other hand, errors of morphological analysis occur in a local part in a compound, and character-based HMM can avoid such errors.

Accordingly, morphological analysis and character-based HMM are complementary to each other. In this paper, we correct word segmentation errors produced by morphological analysis through character-based HMM. The question is how to find errors of morphological analysis. In this paper, we import error patterns of morphological analysis. If word segmentation by morphological analysis includes a prepared error pattern, we compare it with the corresponding part of word segmentation by character-based HMM. If they are different, we correct the former to the latter.

Last we note that we use JUMAN 3.5 as a morphological analysis system.

1. WORD SEGMENTATION BY CHARACTER-BASED HMM

1.1. Character-based HMM

HMM M is defined by six elements: S, Y, A, B, π and F. S is a set of states, Y is a set of output symbols, A is a set of state translation costs, B is a set of output costs, π is a set of initial state costs and F is a set of final states.

The problem of word segmentation is transposed to the problem of assignment of the word boundary sign (1) or not word boundary sign (0) between each two characters. Thus, we set S to be $\{0,1\}$, π and F to be $\{1\}$, and Y to be a set of bi-gram of Japanese character. Moreover, we ignore A. The setting of B constructs the character-based HMM M.

In HMM, we can estimate the sequence of states in which observed output symbols pursue by the Viterbi algorithm. That is, we can estimate whether the word boundary exists between two characters or not. Accordingly, we can segment a compound into words (refer to Figure 1).

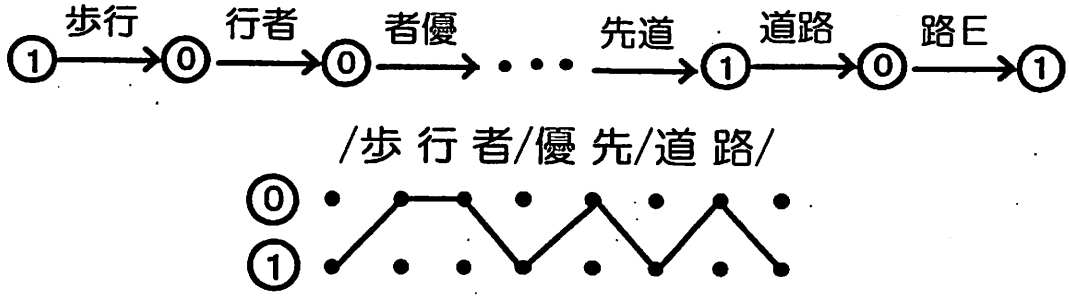

FIGURE 1. Word segmentation through HMM

1.2. Calculation of output costs for characters

Let $B_{ij}(ab)$ be the costs to output the character bi-gram ab with movement from the state i to the state j. In this case, calculation of $B_{11}(ab)$, $B_{10}(ab)$, $B_{01}(ab)$ and $B_{00}(ab)$ for each character bi-gram ab corresponds to the construction of B.

To obtain values of $B_{ij}(ab)$, we conducted morphological analysis for one-year's worth of newspaper articles from Nikkei Shinbun CD-ROM '90, and picked up compounds. For each character in a picked-up compound, we can know whether the word boundary exists on the left (and the right) of that character or not. Let $C(i,j,a)$ be the number of character a such that its left state is i and its right state is j. For example, the compound "自然言語" is segmented into /自然/言語/ by morphological analysis. Thus, $C(1,0,\text{自})$, $C(0,1,\text{然})$, $C(1,0,\text{言})$ and $C(0,1,\text{語})$ are added to 1. We define $B_{ij}(a)$ as follows:

$$B_{ij}(a) = \log_2 \frac{C(i,\bar{j},a)+1}{C(i,j,a)+1}$$

Note that $\bar{j}$ is the state of not j, that is, $\bar{j}=0$ if $j=1$ and $\bar{j}=1$ if $j=0$.

Next we define $C(i,j,ab)$. This means the number of character a such that its left state is i and its right state is j and the next character of a is b. $C(i,j,ab)$ is obtained through compounds picked up above. Moreover, let $B'_{ij}(ab)$ be the costs to output the character a with movement from the state i to the state j, where the next character of a is b. We define

it as follows:

$$B'_{ij}(ab) = \log_2 \frac{C(i,\bar{j},ab)+1}{C(i,j,ab)+1}$$

Last we define $B_{ij}(ab)$ by linear combination of $B_{ij}(a)$ and $B'_{ij}(ab)$ as follows:

$$B_{ij}(ab) = \alpha \cdot B_{ij}(a) + (1-\alpha) \cdot B'_{ij}(ab)$$

In this paper, we set α to be 0.3.

2. CORRECTION OF WORD SEGMENTATION ERRORS BY CHARACTER-BASED HMM

2.1. Errors of character-based HMM

Most errors of character-based HMM are caused by words composed of many characters, because character-based HMM judges whether the word boundary exists between two characters or not from the relation between these two characters and one or two characters surrounding them. For example, "1次産品共通基金" is one word in the JUMAN dictionary. It is not a question of whether that character sequence is one word or not. In the required application, if that character sequence must be handled as one word and is registered as one word in the dictionary, we should judge it to be one word.

The correct word segmentation of this character sequence "1次産品共通基金" is /1次産品共通基金/. On the other hand, the correct word segmentation of the character sequence "1次産品" is /1/次/産品/. That is, we cannot judge whether the word boundary between "1" and "次" exists or not, by using only the relation between these two characters and one or two characters surrounding them. Moreover, we cannot judge it even by using the relation between "1" and "次産" and relation between "1" and "次産品"

However, it is easy for morphological analysis to correctly segment "1次産品共通基金" and "1次産品" into words.

2.2. Errors of morphological analysis

Most errors of morphological analysis are caused by unknown words. Most Japanese words are composed of two or three characters. Thus, unknown word composed of two characters "○○" is segmented as /○/○/ and unknown word composed of three characters "○○○" as /○/○/○/ or /○/○○/.

Besides the cases of unknown words, word segmentation errors such as /○/○○～ for /○○/○～ is typical in JUMAN.

We can find a common characteristic among the above patterns, that is, the length of the first word in segmented words is one. For that reason, the part including a word composed of one character may be incorrect segmentation.

On the other hand, character-based HMM is effective for such local segmentation.

2.3. Complementary use to each other

As we have said, the respective weaknesses of morphological analysis and character-based HMM compensate each other.

In this paper, we propose a method in which the morphological analysis and character-based HMM are complementary to each other, in order to segment a compound into words. First we conduct word segmentation by the morphological analysis and character-based HMM

simultaneously side by side, and then compare the two results. Basically we accept the result from morphological analysis. However, the part including the pattern /○/○~/ is modified to the corresponding part /○○~/ of the result of character-based HMM.

We note that the length of /○/○~/ and the length of /○○~/ are the same, and the positions of word boundary mark "/" in /○/○~/ and in /○○~/ match only in the first position and in the last position. The underlined parts in Table 1 are examples. In this paper, we name this pattern *P*0.

TABLE 1. Examples of modification

Morphological analysis	HMM
/鈴木/健/四郎/	/鈴木/健四郎/
/永島/帝/二/さん/	/永島/帝二/さん/
/河/口利/加さん/	/河口/利加/さん/
/東/大卒/	/東大/卒/

3. EXPERIMENTS

We extracted 8,543 kinds of compounds from newspaper articles of Mainichi Shinbun CD-ROM '94 in order of appearance. We conducted word segmentation by the morphological analysis and character-based HMM simultaneously side by side. Two word segmentations for 7,760 kinds of compounds (90.8%) were the same, but two word segmentations for 783 kinds of compounds (9.2%) were different. The number of word segmentations produced by morphological analysis, which includes the pattern *P*0, was 212 (2.5%). For these 212 kinds of word segmentation, we modified them by word segmentation produced by character-based HMM. We show a part of them in Table 2:

TABLE 2. Correction of segmentation

morphological analysis	HMM
/延/岡市/	/延岡/市/
/奥谷/喬/司/	/奥谷/喬司/
/追/加点/	/追加/点/
/病/人食/	/病人/食/
/島/内/監督/	/島内/監督/
/若/田光/一/さん/	/若田/光一/さん/
/若/貴兄/弟/	/若貴/兄弟/
...	...
/小泉/順/一郎/郵政相/	/小泉/順一郎/郵政相/
/織田/大/次郎/店長/	/織田/大次郎/店長/

We checked whether the modified 212 kinds of word segmentation are correct or not. As a result, 178 kinds of word segmentation (84.0%) were correct, but 34 kinds of word segmentation (16.0%) were incorrect. However, for 15 kinds of word segmentation in these 34 modifications, the original word segmentation produced by morphological analysis were incorrect, too. Hence, 19 kinds of modification were adversely affected.

4. REMARKS

4.1. Ratio of errors covered with the pattern

In this paper, we are concerned with only errors with the pattern $P0$. Errors with other patterns exist, so the above experiments did not show by how much the precision of word segmentation is improved.

Here we randomly picked up 783 kinds of compounds from 8,543 kinds of compounds used in the above experiments, conducted morphological analysis for them, and checked whether they were correct or not. As a result, we found 28 kinds of word segmentation errors, and 26 kinds of them had the pattern $P0$. Our method can correct 84.0% of errors with the pattern $P0$, so we estimate that our method can correct 78.0% ($= 0.928 * 0.840$) of errors produced by morphological analysis.

4.2. Word segmentation for compounds in other domains

In the experiment, we used compounds from newspaper articles. Training data is also obtained from newspaper articles. Therefore the evaluation for our method may be optimistic.

Here we confirm the availability of our method for compounds in other domains. We randomly picked up technical terms with the compound form from three domains: the economics, the information science, and the medical science. Next, we applied our method to these compounds. The result was as follows:

TABLE 3. Result for compounds in other domains

domain	compounds	modification	valid	not valid	both error
economics	313	13	9 (69.2%)	3	1
information science	258	21	14 (66.7%)	5	2
medical science	345	113	39 (34.5%)	37	37

This table shows that our method is valid for compounds in domains of the economics and the information science, but is not valid for compounds in the domain of the medical science. We guess that it is because terms in domains of the economics and the information science are frequently appeared in newspaper articles, but terms in the domain of the medical science are not.

That is, our method is not so available for compounds in the very different domain from training data. However, Table 3 shows that the pattern $P0$ is valid.

4.3. Assignment of part of speech

Our method has the defect that the modified segmentation lacks a part of speech. We can partially cope with this problem by generating all possible word segmentations through JUMAN. For example, JUMAN produces 4 possible word segmentations for "東大卒". The segmentation /東大/卒/ in them is only one /東大 (organization name)/卒 (normal noun)/.

By using this information, we can assign parts of speech to the modified segmentation.

4.4. Relevant research

We used the character-based HMM not requiring a dictionary to cope with unknown words. Other methods not requiring a dictionary involve measuring the strength of the

connection between characters. In order to measure the strength of the connection, the use of mutual information (Sproat 1990) and the likelihood ratio test of hypotheses (Kageura 1997) have been proposed. However, these methods have also the same problem as the character based-HMM, that is, cannot recognize a word composed of many characters as one word.

We obtained parameters of character-based HMM by using word segmentation produced by morphological analysis, so our method is classified as a supervised learning method. Hence, our method cannot be exported to a new domain which includes many unknown words. In this case, an unsupervised learning method (Nagata 1997; Sproat 1996; Luo 1996 etc.) is required. This will be our future work.

5. CONCLUSION

In order to segment a compound into words, this paper proposed a method in which morphological analysis and character-based HMM are complementary to each other.

Character-based HMM cannot recognize a word composed of many characters as one word, while errors of morphological analysis are local and have a particular pattern. The part of segmentation with this pattern is modified by segmentation produced by character-based HMM.

In experiments, we applied our method to 8,543 kinds of compounds. As a result, our method corrected 178 kinds of incorrect word segmentation produced by morphological analysis, but 19 kinds of modification were adversely affected. From the ratio of errors covered with the used pattern, we estimate our method can correct about three fourths (78.0%) of errors which morphological analysis will produce.

REFERENCES

KAGEURA, K. 1997. Mojitani no bigram syakudo ni motozuku fukugougo kanjiretu no tangokiri syuhou (in Japanese). In *The 3rd Annual Meeting of the Association for Natural Language Processing* : 477–480.

LUO, X., and S. ROUKOUS. 1996. An Iterative Algorithm to Build Chinese Language Models. In *The 34th Annual Meeting of the Association for Computational Linguistics (ACL-96)* : 139–143.

NAGATA, M. 1997. A Self-organizing Japanese Word Segmentation using Heuristic Word Identification and Re-estimation (in Japanese). Technical Report NL-121-2, IPSJ SIG Notes of NL.

ODA, H., and K. KITA. 1998. Japanese Word Segmentation by a PPM^* Model (in Japanese). Technical Report NL-128-2, IPSJ SIG Notes of NL.

SPROAT, R., and C. SHIH. 1990. A statistical method for finding word boundaries in Chinese text. *Computer Processing of Chinese and Oriental Language*, 4 : 336–351.

SPROAT, R., C. SHIH., W. GALE., and N. Chang. 1996. A Stochastic Finite-State Word-Segmentation Algorithm for Chinese. *Computational Linguistics*, Vol. 22, No. 3 : 377–404.

TUSJI, K., and K. KAGEURA. 1997. An HMM-based Method for Segmenting Japanese Terms and Keywords based on Domain-Specific Bilingual Corpora. In *The 4th Natural Language Processing Pacific Rim Symposium* : 557–560.

YAMAMOTO, M., and M. MASUYAMA. 1997. Hinshi Kugiri jouhou wo fukumu kautyoumoji no rensakakuritu wo motiita nihongo keitaisokaiseki (in Japanese). In *The 3rd Annual Meeting of the Association for Natural Language Processing* : 421–424.