Revision of Morphological Analysis Errors Through the Person Name Construction Model

Hiroyuki Shinnou

Ibaraki University
Dept. of Systems Engineering
Nakanarusawa, 4-12-1
Hitachi, Ibaraki, 316-8511, Japan
shinnou@lily.dse.ibaraki.ac.jp

Abstract. In this paper, we present the method to automatically revise morphological analysis errors caused by unregistered person names. In order to detect and revise their errors, we propose the Person Name Construction Model for kanji characters composing Japanese names. Our method has the advantage of not using context information, like a suffix, to recognize person names, thus making our method a useful one. Through the experiment, we show that our proposed model is effective.

1 Introduction

It is clear that morphological analysis is an important module for an NLP system like the MT system. One problem in the morphological analysis is word segmentation errors caused by unregistered words. Most unregistered words are proper nouns, like place names, organization names, and person names. In this paper, we focus on person names, and propose the Person Name Construction Model to correct morphological analysis errors caused by unregistered person names. This model gives a score to the given word sequence. This score indicates the degree to which the given word sequence appears to be a person's name. By the score, we can extract the name from the morphological analysis result. If the extracted name is not consistent with the morphological analysis result, we revise the result to take the extracted person name into account.

The Person Name Construction Model is based on the heuristic that a person's name is composed of kanji characters which are placed in the first position, the middle position and the last position of the name. For example, in the case of the family name, kanji characters frequently used in the first position are "中", "松" and "長". And in the middle position, they are "谷", "々", "曽". And in the last position, they are "田", "藤", "井". Our proposed model deduces that the character sequences "中谷田", "長藤" and "松曽井", which are a combination of their characters, have the appearance of being names[1]. However, this model tends to judge the given character sequence to be a person's name. So, in order

[1] These character sequences are not registered in the dictionary. We don't know whether these character sequences are real person names. However, most Japanese agree that these character sequences seem to be family names.

to remove non-names from the extraction, we also use the heuristic based on the morphological analysis error patterns caused by unregistered person names.

A feature of our proposed model makes no use of contextual clues. Strategies to recognize unregistered words are divided into two types. The one type uses contextual clues, like a suffix (ex. "氏 (Mr.)", "さん (Ms..)"), a prefix (ex. "故 (the late)", "長女 (the first-born daughter)"), the initial phrasing(ex. "社長 (the president)", "大統領 (the President")), a verb (ex. "逮捕される (be arrested)", "殺される (be killed)") and so on in order to recognize unregistered words[3, 8]. Another type uses only clues in the given word sequence, and doesn't use information out of the given word sequence. The former is powerful, and currently the automatic acquisition of such contextual clues is being researched[6, 1, 5]. However we often have the situation without contextual clues. Thus the former strategy needs to have the latter strategy module. For example, in the case of the phrase "～社長 (the president ～)", the "～" part often includes a name. Thus, the phrase "～社長" is a contextual clue to recognize person names. However, the "～" words in this phrase do not always include a person's name. Therefore from only information in the "～" sequences, we must judge whether it includes a person's name or not. Our proposed model is useful in doing this, and can be applied to all sorts of former strategies.

Last we experimented using a small sampling. For morphological analysis errors caused by unregistered person names, our system revised them with 63.8% precision and 72.5% recall. Investigating our system failures, we found most failures acceptable and reasonable. So our proposed model was shown to be useful and effective for the recognition of unregistered person names.

2 Extraction of person names and revision of morphological analysis errors

2.1 Basic procedures

First, we pick out kanji word sequences for doing a morphological analysis of a sentence. Here, we define the term "kanji word" as words composed of kanji characters. For example, for the following sentence (1), we get sentence (2) as the result of a morphological analysis, and we pick out the three kanji word sequences shown in (3).

(1) あの千葉大学の学生が鈴木健四郎社長です
(That student going to Chiba university is the president Suzuki Kensirou.)
(2) /あの/千葉/大学/の/学生/が/鈴木/健/四郎/社長/です/
(3) /千葉/大学/, /学生/, /鈴木/健/四郎/社長/
(Chiba university, student, the president Suzuki Kensirou)

A name is extracted from each kanji word sequence if the sequence contains a person's name. If the extracted name is not consistent with the morphological analysis result, we correct the morphological analysis result to account for the extracted name.

Table 1. Person names extracted form kanji word sequences

kanji word sequence	extracted person name
/千葉/大学/	/千葉/ (last name) ...(4)
/学生/	nothing
/鈴木/健/四郎/社長/	/鈴木/健四郎/ (last name/first name)...(5)

For the above example, we have extracted the names shown in Table 1.

From the kanji sequence "/千葉/大学/", we extract the name "/千葉/" as the last name (c.f. (4)). This segmentation is consistent with the morphological analysis result, so we don't revise it. On the other hand, the sequence "/健四郎/", extracted as the first name from the kanji sequence "/鈴木/健/四郎/社長/", is not consistent with the morphological analysis result, in which "健四郎" is segmented into "/健/" and "/四郎/". Therefore, we revise the morphological analysis result to the sequence "/健四郎/".

Next, we describe the procedure to extract the person's name from the kanji sequence. First we extract kanji word subsequences as a part of the given kanji word sequences, and we give each kanji word subsequence a score which indicates the degree to which the given kanji word subsequence appears to be a person's name. Next, we identify the kanji word subsequence as a name if its maximum score goes over a threshold value. The output is the kanji word subsequence recognized as the person's name and classified by type (i.e. last name, first name, or their combination).

Take the case of the kanji word sequence "/鈴木/健/四郎/社長/". We extract kanji word subsequences from its sequence, and get the score for each kanji word subsequence as shown in Table 2. We output the phrase "/鈴木/健四郎/" with the maximum score.

2.2 Person Name Construction Model

Our system computes a score which indicates the degree to which the given word sequence appears to be a person's name. In order to compute the score, we propose the Person Name Construction Model.

Japanese names consist of a last name and first name. Last names can be divided into three character parts: the first position character(LFC), the middle position character(LMC) and the last position character(LLC). For instance, the last name "中曽根" has following the three character parts.

$$LFC = \text{“中”},\ LMC = \text{“曽”},\ \text{and}\ LLC = \text{“根”}.$$

In the case of the last name "鈴木", the character parts are:

$$LFC = \text{“鈴”}, LMC = \text{“”},\ \text{and}\ LLC = \text{“木”}.$$

Table 2. Score for the kanji word subsequence

kanji word subsequence	score	extracted name
/鈴木/健/四郎/社長/	0	nothing
/鈴木/健/四郎/	338970480	/鈴木/健四郎/ (last name/first name)
/鈴木/健/	4014368	/鈴木/健/ (last name/first name)
/鈴木/	5296	/鈴木/ (last name)
/健/四郎/社長/	0	nothing
/健/四郎/	43167	/健四郎/ (first name)
/健/	758	/健/ (first name)
/四郎/社長/	0	nothing
/四郎/	5906	/健四郎/ (first name)
/社長/	0	nothing

In the same way, first names can divided into three character parts: the first position character(FFC), the middle position character(FMC) and the last position character(FLC).

Our model assumes that any kanji character "a" has a score which indicates how often the character "a" is used as an LFC. Also the character "a" has scores for LMC and LLC. We define $S_{lfc}(a)$ to be the LFC score for a character "a". We define $S_{lmc}(a)$ and $S_{llc}(a)$ similarly. By the following expression, we define the score $S_l(\alpha)$, which indicates the degree to which a character sequence $\alpha = a_1a_2a_3\cdots a_n$ appears, to be a last name.

$$S_l(\alpha) = \frac{S_{lfc}(a_1) + \sum_{i=2}^{n-1} S_{lmc}(a_i) + S_{llc}(a_n)}{n}$$

In the same way, in the following expression, we define the score $S_f(\beta)$, which indicates the degree to which a character sequence $\beta = b_1b_2b_3\cdots b_n$ appears, to be a first name.

$$S_f(\beta) = \frac{S_{ffc}(b_1) + \sum_{i=2}^{n-1} S_{fmc}(b_i) * S_{fec}(b_n)}{n}$$

Finally, in the following expression, we define a score indicating the degree to which a string α appears to be a last name and a string β a first name.

$$S_l(\alpha) * S_f(\beta)$$

If the length of the character sequence is over 2, we can calculate the score for the character sequence. If the length of the character sequence is 1, i.e. the character sequence is $\alpha = a_1$, we define the scores as follows:

$$S_l(\alpha) = S_{l1}(a_1)$$

$$S_f(\alpha) = S_{f1}(a_1)$$

We will define $S_{l1}(a_1)$ and $S_{l1}(a_1)$ later.

When we are given the kanji word subsequence $P = w_1 w_2 \cdots w_m$, we regard it as the character sequence $P = a_1 a_2 \cdots a_n$. Next we compute each score of $S_l(P)$,$S_f(P)$ and $S_l(a_1 a_2 \cdots a_i) * S_f(a_{i+1} a_{i+2} \cdots a_m)$, and output one with the maximum score.

Finally, we must explain how to construct scores of $S_{lfc}(a_1)$ and so on. In this paper, we used one-year-old newspaper articles as the training corpus. First, we segmented words by morphological analysis for the training corpus. We identified person's names as a result of the morphological analysis, and made a frequency table (T1) for these names. And then we picked out person's names from the dictionary used for morphological analysis and made a frequency table (T2) for these names. T2 always has a frequency of 1. Next we merged T1 and T2, and divided it into a frequency table (TL) for last names and a frequency table (TF) for first names. Further, we divided TL into a frequency table (TL1) for names of the length 1 and a frequency table (TL2) for names of length 2 or over. Similarly, we got TF1 and TF2. Next if the frequency of the name $\alpha = a_1 a_2 a_3 \cdots a_n (n > 1)$ in TL2 is f, we add the value f to the $S_{lfc}(a_1)$,$S_{lmc}(a_2)$, $S_{lmc}(a_3)$,$\cdots$, $S_{lmc}(a_{n-1})$ and $S_{fec}(a_n)$. We repeated this procedure for all names in TL2. As a result we arrived at scores $S_{lfc}(a)$,$S_{lmc}(a)$ and $S_{llc}(a)$. And we also got scores $S_{ffc}(a)$,$S_{fmc}(a)$ and $S_{fec}(a)$ in this same way.

We defined $S_{l1}(a)$ and $S_{f1}(a)$ to be the frequency of the last name "a" and the first name "a", so these scores can be defined in TL1 and TF1.

Lastly, we explain the case that $S_{*c}(a)$ or $S_{*1}(a)$ is equal to zero. In that case, basically $S_l(\alpha)$ or $S_f(\alpha)$ is defined to be zero. However, if the character sequence α has the following form:

last name + first name,

we used 10 % of $S_l(\alpha)$ as $S_l(\alpha)$, and 10 % of $S_f(\alpha)$ as $S_f(\alpha)$,

2.3 Use of morphological analysis result

The Person Name Construction Model tends to extract too many names from kanji word sequences. This occurs because this model measures the appearance of the person name, although appearance is a weak indication of a person's name. Therefore, it is difficult to judge by only these characteristics whether or not the kanji word sequence is a person's name.

In this paper, we use the result of morphological analysis, together with the Person Name Construction Model. First, we have applied the following heuristics.

H0 Morphological analysis error caused by the unregistered person name includes the kanji word whose length is 1.

For example, the first name "健四郎" is segmented into "/健/" and "/四郎/", but this segmentation is wrong. This morphological analysis error includes the kanji word "/健/" whose length is 1. Most Japanese names have a length of 1,2 or 3. So, if a morphological analysis has incorrectly segmented a part of a person's name, it is clear that a kanji word with length 1 is included.

By using the heuristics H0 and the dictionary, we can judge that a kanji word sequence isn't a person's name. It should be noted that the heuristics H0 does not help us to judge whether a kanji word sequence is a name. If we can judge that the given kanji word sequence isn't a name, the score is zero, and if we cannot judge, the score is obtained by using the Person Name Construction Model.

Next, for the kanji word sequence which includes a kanji word with length 1, we use the following heuristics.

H1 If a morphological analysis error caused by the unregistered person name includes the kanji word whose length is 2, this kanji word is a person's name.

In the above example, the morphological analysis error (segmentation into "/健/" and "/四郎/") for the first name "健四郎" includes the kanji word "/四郎/" with length 2, and this word is a person's name. The heuristics H1 seems tenuous. However we confirmed it to be effective by the following experiment. First we picked person names with length 3 from the dictionary. If the picked word has a character string of $k_1k_2k_3$, we made the character strings k_1k_2 and k_2k_3, and checked whether k_1k_2 or k_2k_3 is a person's name. 78.0% of the picked names k_1k_2 or k_2k_3 resulted as person names. This experiment shows that the heuristics H1 is effective.

By using the heuristics H1, we can judge that a kanji word sequence is not a person's name. Again note that the heuristics H1 cannot judge that a kanji word sequence is a person's name. If we can judge that the kanji word sequence isn't a person's name, the score is zero, and if we cannot judge it, the score is obtained by the Person Name Construction Model.

Lastly we use the following heuristics:

H2 "numeral word + suffix word" is not a person's name.

This pattern appears frequently. The kanji word sequence "/千/円/" is an example of this. We assume that these kanji word sequences are not person names.

2.4 Collection of revision error

Even if we use the proposed model and heuristics H0,H1 and H2, some kanji word sequences are judged wrongly as person names. However the frequency of these wrong revision patterns is low, and we gathered frequent revision errors to avoid these errors.

First, we did morphological analysis on a part of the training corpus [2]. Next we revised morphological analysis errors with our system. We collected revised person names, and made a frequency table for the names. Because the frequency of general person names is low, names with high frequency are regarded as wrong

[2] 10% of training corpus

revisions. Through these experiments, we registered the following 10 phrases as non-names.

"日米", "対米", "対中", "国間", "花博", "各行", "一極", "信金", "安門", "日債"

3 Experiment

To confirm that our proposed model is useful and effective, we picked 1,095 sentences from the beginning of newspaper articles[3], and experimented with them. We did a morphological analysis of these sentences using the JUMAN system [4]. Investigating the results of the morphological analysis on them, we found 51 errors (42 kinds) caused by unregistered person names. Our system revised 58 phrases (41 kinds) that resulted from the morphological analysis. A correction was made on 37 phrases (28 kinds). This result shows that the precision rate was 63.8% and the recall rate was 72.5%.

The corrections are shown in Table 3.

Table 3. Right revisions

kanji word sequence	correction
/吉村/午/良/知事/	/吉村/午良/ (last name/first name)
/橋本/大/二郎/知事/	/橋本/大二郎/ (last name/first name)
/小/渕/恵三/自民党/副総裁/	/小渕/恵三/ (last name/first name)
/木/見/金治郎/門下/	/木見/金治郎/ (last name/first name)
/米/長/	/米長/ (last name)
/沢/近/	/沢近/ (last name)
/岩/國/哲人/	/岩國/哲人/ (last name/first name)
...	...

Our system could not detect 14 morphological analysis errors (13 kinds) caused by unregistered person names. We have classified the reasons for this into the following 4 types.

1. Segmentation of a registered word is wrong (2 errors, 2 kinds).
   These two kanji word sequences were segmented as follows:
   - /井上/雅/晶代/表/ (Right segmentation is /井上/雅晶/代表/)
   - /３１/日田/嶌/徳弘/ (Right segmentation is /３１/日/田嶌/徳弘)

[3] Mainichi Shinbun '95 CD-ROM.
[4] JUMAN is a standard Japanese morphological analysis system

The registered words ("/代表/" and "/日/") were also wrongly segmented like those above. Because our system assumes that there are none of these types of errors, our system cannot extract the name or revise this type of error.

2. Person's name is foreign (7 errors, 2 kinds).
   For example, "/鍾/仕/梅/" and "/王/文/煥/" are morphological analysis errors. But these name are Chinese or Korean names.
   Because our proposed model is based on heuristics founded on Japanese person name, our model can basically not detect this type of error.
3. Person's name is old (3 errors, 3 kinds).
   The three names are "/大橋/宗/桂/", "/本因坊/算/砂/" and "/算/砂/". Because we used current newspapers as the training corpus, it is difficult to devise the model parameters for old Japanese names. These name are not covered by our model.
4. Person's name is very rare (2 errors, 2 kinds).
   These name are "/楠/部/彌/弍/" and "/田原/護/立/". Our model must revise these errors, but was not able to do this.

Only the 4th error type has been unsatisfactory in our proposed model, so it is reasonable to assume that our proposed model is effective.

Next we classify revision errors (19 errors, 16 kinds) into 4 types as follows.

- The detection that the given kanji word sequence includes a person's name, is successful, but revision fails (4 errors, 3 kinds).
  In the case of morphological analysis error "/楠/部/彌/弍/", we revised "/楠/部/" to "/楠部/" (last name), but this is wrong. The right revision is "/楠部/彌弍/" (last name/first name). In this case, the system has successfully detected that the kanji word sequence "/楠/部/彌/弍/" includes a person's name.
- The unregistered proper noun which is not a person's name is revised (9 errors, 7 kinds).
  In the case of morphological analysis error "/星/島/日報/", we revised "/星/島/" to "/星島/" (last name). The kanji word sequence "/星/島/" is an unregistered proper noun, and the right segmentation is "/星島/". So our revision is effective, but the word "/星島/" is not a person's name.
- The unregistered proper noun which is not a person name is detected, but the revision fails (5 errors, 5 kinds).
  In the case of morphological analysis error "/油/麻/地/", we revised "/麻/地/" to "/麻地/" (last name). The kanji word sequence "/油/麻/地/" is an unregistered proper noun, and the right segmentation is "/油麻地/". We successfully detected that this kanji word sequence includes unregister words, but failed to judge the unregistered word is a person's name, and the segmentation for it failed.
- Results of morphological analysis were correctly revised (1 error, 1 kind).
  For example, we correctly revised the segmentation "/東/口/" to "/東口/" (last name).

Generally, we cannot judge without contextual information whether a proper noun is a person's name or not. Therefore, we cannot avoid the 2nd type error. The recognition of an unregistered word is useful in NLP systems. As for our system, only the 4th error is regarded as a failure.

In conclusion, We should note that our proposed model is useful and effective.

4 Remarks

The aim of our system is the automatic revision of morphological analysis errors caused by unregistered person names. However, our system can be used as the extraction system for person names. In recent years, the information extraction systems have been actively researched[4]. In these systems, it is important to correctly extract person names from texts[7]. Our system is useful in this aspect. The problems of extracting person names are classified into the following 3 types. These type phenomena make it difficult to extract names.

1. Morphological analysis errors cause by unregistered words.
   For example, the right segmentation for the character sequence "鈴木健四郎" is "/鈴木/健四郎/", but a morphological analysis wrongly segments it as "/鈴木/健/四郎/", because the first name "健四郎" is unregistered.
2. Assignment of part of speech fails.
   For example, a morphological analysis correctly segments "細川正" as "/細川/正/", but the part of speech for "細川" is assigned as a general noun. This is wrong. The part of speech for "細川" is the person's name.
3. The word is correctly judged as a person's name upon morphological analysis, but the word is not a person's name in the context.
   For example, a morphological analysis correctly segments "松下塾" as "/松下/塾/", and the part of speech for "松下" is correctly assigned as a person's name. However, in information extraction, the word "松下" should not be extracted as a person's name, because the phrase "/松下/塾/" is the organization name.

Our system can be useful in solving the first problem. The 2nd and 3rd problems cannot be solved without contextual information. Contextual information is also useful for the 1st problem. However, as mentioned in the introduction, even the method using contextual information needs to judge whether the given word sequence is a person's name or not. And our model can be used together with all methods using contextual information. The improvement of the module, which judges whether the given word sequence is a person's name or not, directly improves the extraction system of person names.

A fault of our system is that scores are defined by heuristic method. We should define scores by probability. However, it is unclear how to make the score correspond to the probability, and how to determine probabilities. A definition of the score based on frequency like our system is simple, and works well. Consideration of this aspect will improve our system.

Our model deals with Japanese names and not foreign names. However, foreign names expressed by kanji characters are almost always Chinese names or Korean names. There are a limited number of last names of Chinese and Korean, and there is a heuristic that the length of the last name is 1 and the length of the first name is 2[2]. We believe that it is easy to recognize unregistered Chinese names and Korean names in Japanese texts.

5 Conclusion

In this paper, we presented the method to automatically revise morphological analysis errors caused by unregistered person names. The main part of our method is the module to give the word sequence a score which indicates the degree to which it appears a person's name. To implement this module, we proposed the Person Name Construction Model which applies the heuristic rule on kanji characters composing Japanese names. Through the experiment, we have shown that our proposed model is effective and useful. The problem of our revision system is how to define scores. For this problem, the import of probability may be effective. This is our future task.

Acknowledgments

We used Nikkei Shibun CD-ROM '90 and Mainichi Shibun CD-ROM '95 as the corpus. The Nihon Keizai Shinbun company and the Mainichi Shinbun company gave us permission of use of their collections. We appreciate the assistance granted by both companies.

References

1. Bikel,D., Miller,S., Schwartz,R. and Weischedel,R. : "Nymble: a High-Performance Learning Name-finder", Proc. of ANLP-97, pp. 194–201 (1997).
2. Chen,H-H. and Lee,J-C. : "Identification and Classification of Proper Nouns in Chinese Texts", Proc. of COLING-96, pp. 418–424 (1996).
3. Chen,K-J. and Liu, S-H. : "Word Identification for Mandarin Chinese Sentences", Proc. of COLING-92, pp. 101–107 (1992).
4. Grishman,R. and Sundheim,B. : "Message Understanding Conference-6: A Brief History", Proc. of COLING-96, pp. 446–471 (1996).
5. Sekine,S., Grishman,R. and Shinnou,H. : "A Decision Tree Method for Finding and Classifying Names in Japanese Texts", Proc. of WVLC-6, to appear (1998).
6. Strzalkowski,T. and Wang,J. : "A Self-Learning Universal Concept Spotter", Proc. of COLING-96, pp. 931–936 (1996).
7. Wakao,T., Gaizuska,R. and Wilks,Y. : "Evaluation of an Algorithm for the Recognition and Classification of Proper Names", Proc. of COLING-96, pp. 418–424 (1996).
8. Wang,L-J., Li,W-C. and Chang, C-H. : "Recognizing Unregistered Names for Mandarin Word Identification", Proc. of COLING-92, pp. 1239–1243 (1992).