Computational Linguistics and Chinese Language Processing
Vol. 17, No. 3, September 2012, pp. 87-108

Strategies of Processing Japanese Names and Character Variants in Traditional Chinese Text

Chuan-Jie Lin[*], Jia-Cheng Zhan[*], Yen-Heng Chen[*], and

Chien-Wei Pao[*]

Abstract

This paper proposes an approach to identify word candidates that are not Traditional Chinese, including Japanese names (written in Japanese Kanji or Traditional Chinese characters) and word variants, when doing word segmentation on Traditional Chinese text. When handling personal names, a probability model concerning formats of names is introduced. We also propose a method to map Japanese Kanji into the corresponding Traditional Chinese characters. The same method can also be used to detect words written in character variants. After integrating generation rules for various types of special words, as well as their probability models, the F-measure of our word segmentation system rises from 94.16% to 96.06%. Another experiment shows that 83.18% of the 862 Japanese names in a set of 109 human-annotated documents can be successfully detected.

Keywords: Semantic Chinese Word Segmentation, Japanese Name Identification, Character Variants.

1. Introduction

Word segmentation is an indispensable technique in Chinese NLP. Nevertheless, the processing of Japanese names and Chinese word variants has been studied rarely. At the time when Traditional Chinese text was mostly encoded in BIG5, writers often transcribed a Japanese person’s name into its equivalent Traditional Chinese characters, such as the name “滝沢秀明” (Hideaki Takizawa) in Japanese becoming “瀧澤秀明” in Traditional Chinese. After Unicode was widely adopted, we also could see names written in original Japanese Kanji in Traditional Chinese text. Another issue is how different regions may write a character

[*] Department of Computer Science and Engineering, National Taiwan Ocean University
No 2, Pei-Ning Road, Keelung, 20224 Taiwan
E-mail: cjlin@ntou.edu.tw; jjt@cyber.ntou.edu.tw; {M97570019, M97570020}@ntou.edu.tw

in a different shape. For example, the Traditional Chinese character 圖 (picture) is written as 图 in Simplified Chinese and 図 in Japanese. How these character variants impact Chinese text processing has been mentioned rarely in earlier studies; thus, it has become our interest.

Chinese word segmentation has been studied for a long time. Many recent word segmentation systems have been rule-based or probabilistic. The most common rules are longest-word-first or least-segmentation-first. The probability models are often built in Markov's unigram or bigram models, such as in Peng and Chang (1993). Word candidate sets are often vocabulary in a dictionary or a lexicon collected from a large corpus. Some systems also propose possible candidates by morphological rules (Gao *et al.*, 2003), such as NOUN+“們” (plural form of a noun) as a legal word (*e.g*. “學生們,” students, and “家長們,” parents). Wu and Jiang (1998) even integrated a syntactic parser in their word segmentation system.

In addition to word segmentation ambiguity, the out-of-vocabulary problem is another important issue. Unknown words include rare words (*e.g*. “蔓售,” for sale); technical terms (*e.g*. “三聚氰胺,” Melamine, a chemical compound); newly invented terms (Chien, 1997) (*e.g*. “新流感,” Swine flu); and named entities, such as personal and location names. NE recognition is an important related technique (Sun *et al.*, 2003). In recent times, machine learning approaches have been the focus of papers on Chinese segmentation, such as using SVM (Lu, 2007) or CRF (Zhao *et al.*, 2006; Shi & Wang, 2007).

There have been fewer studies focused on handling words that are not Traditional Chinese words in Traditional Chinese text. The most relevant work is discussion of the impact of the different Chinese vocabulary used in different areas on word segmentation systems. These experiments have been designed to train a system with a Traditional Chinese corpus but test on a Simplified Chinese test set or to increase the robustness of a system using a lexicon expanded by adding new terms in different areas (Lo, 2008).

The main problem in this paper is defined as follows. When a word that is not Traditional Chinese appears in a Traditional Chinese document, such as the Japanese name “滝沢秀明” (written in Japanese Kanji) or “瀧澤秀明” (written in its equivalent Traditional Chinese), word variants (*e.g*. “裡面” vs. “裏面”), and words written in Simplified Chinese, all of these words can be detected and become word segmentation candidates. This paper is constructed as follows. Section 2 introduces the basic architecture of our word segmentation system. Section 3 explains the Chinese and Japanese name processing modules. Section 4 talks about the character-variant clusters with a corresponding Traditional Chinese character. Section 5 delivers the experimental results and discussion, and Section 6 concludes this paper.

2. Word Segmentation Strategy

This paper focuses on approaches to handling words that are not Traditional Chinese during word segmentation. We first constructed a basic bigram model word segmentation system. We did not build a complicated system because its purpose is only for observing the effect of applying different handling approaches for words that are not written in Traditional Chinese on the performance of word segmentation. Word candidates were identified by searching the lexicon or applying detection rules for special-type words, such as temporal or numerical expressions. Note that identical word candidates may be proposed by different rules or the lexicon. Moreover, if no candidate of any length can be found at a particular position inside the input sentence, the system automatically adds a one-character candidate at that position. Afterward, the probabilities of all of the possible segmentations are calculated according to a bigram model. The highest probable segmentation is proposed as the result.

2.1 Special-Type Word Candidate Generation Rules

As there are many special type words, it is impossible to collect them all in a lexicon. Hence, we manually designed many detection rules to recognize such words in an input sentence. The special types handled in our system include the following: address, date, time, monetary, percentage, fraction, Internet address (IP, URL, e-mail, *etc*.), number, string written in foreign language, and Chinese and Japanese personal name. Numerical digits in the detection rules can be full-sized or Chinese numbers (一,二…壹貳…). Foreign language characters are detected according to the Unicode table; thus, any character sets, such as Korean or Arabic characters, easily can be added into our system. Consequent characters written in the same foreign language are treated as one word, as most languages use the space symbol as the word-segmentation mark.

Since the focus of this paper is not on the correctness of these special rules, only personal name detection rules will be explained in Section 3.

2.2 Bigram Probabilistic Model

After enumerating all possible segmentations, the next step is to calculate their probabilities $P(S)$. There have been many probabilistic models proposed in word segmentation research. Our system is built on Markov's bigram probabilistic model, whose definition is:

$$P(S = w_1 w_2 ... w_N) = P(w_1) \times \prod_{i=2}^{N} P(w_i \mid w_{i-1}) \qquad (1)$$

where $P(w_i)$ is the unigram probability of the word w_i and $P(w_i \mid w_{i-1})$ is the probability that w_i appears after w_{i-1}. In order to avoid the underflow problem, the equation is often calculated in its logistic form:

$$\log P(S = w_1 w_2 \ldots w_N) = \log P(w_1) + \sum_{i=2}^{N} \log P(w_i \mid w_{i-1}) \quad (2)$$

Data sparseness is an apparent problem, *i.e.* most word bigrams have no probability. Our solution is a unigram-back-off strategy. That is, when a bigram $<w_{i-1}, w_i>$ never occurs in a training corpus, its bigram probability $P(w_i \mid w_{i-1})$ is measured by $\alpha P(w_i)$ instead.

When determining the probability of a bigram containing special-type words, the probability is calculated by Eq. 3. Suppose that w_i belongs to a special type T; the equation is defined as:

$$P(w_i \mid w_{i-1}) P(w_{i+1} \mid w_i) = P(T \mid w_{i-1}) \times P(w_{i+1} \mid T) \times P_G(w_i \mid T) \quad (3)$$

where $P(T \mid w_k)$ and $P(w_k \mid T)$ are the special-type bigram probabilities for the type T and a word w_k and where $P_G(w_i \mid T)$ is the generation probability of w_i being in the type T. The generation probabilities are set to 1 for all special types other than the personal names, whose definitions are explained in Section 3.

As the boundaries of some special types, including address, monetary, percentage, fraction, Internet address, number, and foreign language string, are deterministic and unambiguous, their special-type bigram probabilities are all set to be 1, which means that we accept the segmentation directly.

On the other hand, characters for Chinese numbers often appear as a part of a word, such as “一切” (“一,” one; “一切,” all) and “萬一” (both characters are numbers but together mean “if it happens”). Therefore, the number-type bigram probability is trained from a corpus.

Some temporal expressions are unambiguous, such as the date expression “中華民國九十八年六月二十一日” (“June 21 of the 98th year of the R.O.C.”). Their special-type bigram probabilities are set to 1. For ambiguous temporal expressions, such as “三十年” (meaning “the 30th year” or “thirty years”), their special-type bigram probabilities are obtained by training.

Before training a bigram model, words belonging to special types first are identified by detection rules and replaced by labels representing their types so that special-type bigram probabilities can be measured at the same time.

Our special-type bigram probability model is very similar to Gao *et al.* (2003). Nevertheless, they treat all dictionary words as one class and all types of special words as a second class, while we treat different types as different classes.

2.3 Computation Reduction

When an input sentence is too long or too many possible segmentations can be found (sometimes hundreds of thousands), the computation time becomes intractable. In order to reduce the computation load, we use the beam search algorithm to prune some low probability segmentations. The main idea of the algorithm is described as follows.

Let *N* be the number of characters in an input sentence. Declare *N* priority queues (denoted as record[*i*] where *i* = 1~*N*) to record the top *k* segmentations with the highest probability scores covering the first *i* characters. For each word candidate *w* beginning with the (*i*+1)th character whose length is *b*, append the word *w* with every segmentation stored in record[*i*], compute the probability of the new segmentation, and try to insert it into the queue record[*i*+*b*]. If the new segmentation has higher probability than any segmentation stored in the queue record[*i*+*b*], the segmentation with the lowest probability in record[*i*+*b*] is discarded in order to insert this new segmentation.

At the beginning, all priority queues are empty. Start with the first character in the sentence. Recursively perform the steps described in the previous paragraph until all of the word candidates starting with the *N*th character have been considered. In the end, the top 1 segmentation stored in record[*N*] is proposed as the result. The queue size *k* is set to be 20 in our system.

3. Chinese and Japanese Name Processing

In this section, we focus on how to find Japanese names written in Japanese Kanji that appear in a Traditional Chinese article. The method of identifying Japanese names written in corresponding Traditional Chinese characters is discussed in Section 4. As our approach to recognize Japanese personal names is similar to the one to find Chinese names, our Chinese name identification approach is introduced first.

3.1 Chinese Personal Name Identification

A Chinese personal name consists of a surname part and a first name part. A Chinese surname can be one or two syllables (one or two characters) long. In some cases, a person may have two surnames (usually both with one syllable) in his or her name for various reasons. The first name part in a Chinese name is also one or two syllables long. All name formats possibly seen in an article are listed in Table 1, where “SN” denotes “surname,” “FN” as “first name,” and “char” is “character”.

All strings matching these formats are treated as Chinese name candidates, except the format “1-char FN,” in order to prevent proposing every single character as a personal name candidate. The combination of two surnames is also restricted to two 1-syllable surnames, because one rarely sees a 2-syllable surname combined with another surname. We need to build probabilistic models for each character being in every part of a name, as well as a probabilistic model for the personal name formats.

Table 1. Chinese personal name formats (surnames are underlined)

Format	Cases	Examples	Format	Cases	Examples
SN only	1-char SN	Prof. 林	SN+FN	1-char SN+1-char FN	陳登
	2-char SN	Mr. 諸葛		1-char SN+2-char FN	王小明
FN only	1-char FN	慧		Two SNs+1-char FN	張 李娥
	2-char FN	國雄		Two SNs+2-char FN	張 陳素珠
				2-char SN+1-char FN	諸葛亮
				2-char SN+2-char FN	司馬中原

To recognize a Chinese name, first we have to prepare a complete list of Chinese surnames. We collected surnames from the Wikipedia entries "中國姓氏列表"[1] (List of Chinese Surnames) and "複姓"[2] (2-Syllable Surnames), the websites of the Department of Civil Affairs at the Ministry of Interior[3], 中華百家姓[4] (GreatChinese), and 千家姓[5] (Thousand Surnames). 2,471 surnames were collected. As for the first name part, we simply treat all of the Chinese characters as possible first name characters.

The generation probability model of a word being a Chinese name is defined as Eq. 4, where σ is the gender model (male or female), and π is a possible format matching the word w. The name format is represented as π = 'xxxx,' where 's' denotes a 1-syllable surname, 'dd' a 2-syllable surname, and 'n' a character in a first name. For example, the format "two SNs+2-char FN" is represented as π = 'ssnn' and the format "2-char SN+1-char FN" is represented as π = 'ddn'.

$$P_G(w \mid S_{CHname}) = \max_{\sigma,\pi} P_\sigma(w \mid \pi) P_G(\pi \mid S_{CHname}) \tag{4}$$

In Eq. 4, the ***Chinese name generation probability*** $P_\sigma(w|\pi)$ is the probability of a word w being a Chinese name whose format is π and gender is σ. The ***Chinese name format probability*** $P_G(\pi \mid S_{CHname})$ is the probability of the special type S_{CHname} (Chinese personal names) appearing in an article with a format π. The methods of building these probabilistic models are introduced in the following paragraphs.

When computing the Chinese name generation probability $P_\sigma(w|\pi)$, we borrowed the idea from Chen *et al*. (1998), but we assume that the choice of first names is independent of the surname, and the choice of two characters in the first name part is also independent, in order to reduce the complexity. We also assume that the surname is unrelated to the person's gender. Table 2 lists all of the definitions of the Chinese name generation probabilities for

[1] http://zh.wikipedia.org/wiki/中國姓氏列表

[2] http://zh.wikipedia.org/wiki/複姓

[3] http://www.ris.gov.tw/ch4/0940531-2.doc

[4] http://www.greatchinese.com/surname/surname.htm

[5] http://pjoke.com/showxing.php

every format, where LN_{CH} is the set of Chinese surnames and FN_{CH} is the set of characters used in a Chinese first name. A more sophisticated model may be applied but is outside the scope of this paper.

Table 2. Definitions of the Chinese name probabilities for every name format

Format π	Name Generation Probability $P_\sigma(w\|\pi)$	Format Probability
s	$P_G(c_1\|LN_{CH})$	$P_G(\pi=\text{'s'}\|S_{CHname})$
dd	$P_G(c_1c_2\|LN_{CH})$	$P_G(\pi=\text{'dd'}\|S_{CHname})$
sn	$P_G(c_1\|LN_{CH})\times P_\sigma(c_2\|FN_{CH})$	$P_G(\pi=\text{'sn'}\|S_{CHname})$
nn	$P_\sigma(c_1\|FN_{CH})\times P_\sigma(c_2\|FN_{CH})$	$P_G(\pi=\text{'nn'}\|S_{CHname})$
ddn	$P_G(c_1c_2\|LN_{CH})\times P_\sigma(c_3\|FN_{CH})$	$P_G(\pi=\text{'ddn'}\|S_{CHname})$
snn	$P_G(c_1\|LN_{CH})\times P_\sigma(c_2\|FN_{CH})\times P_\sigma(c_3\|FN_{CH})$	$P_G(\pi=\text{'snn'}\|S_{CHname})$
ssn	$P_G(c_1\|LN_{CH})\times P_G(c_2\|LN_{CH})\times P_\sigma(c_3\|FN_{CH})$	$P_G(\pi=\text{'ssn'}\|S_{CHname})$
ddnn	$P_G(c_1c_2\|LN_{CH})\times P_\sigma(c_3\|FN_{CH})\times P_\sigma(c_4\|FN_{CH})$	$P_G(\pi=\text{'ddnn'}\|S_{CHname})$
ssnn	$P_G(c_1\|LN_{CH})\times P_G(c_2\|LN_{CH})\times P_\sigma(c_3\|FN_{CH})\times P_\sigma(c_4\|FN_{CH})$	$P_G(\pi=\text{'ssnn'}\|S_{CHname})$

The generation probability models for surnames and first name characters, $P_G(c_i|LN_{CH})$, $P_G(c_ic_{i+1}|LN_{CH})$ and $P_\sigma(c_j|FN_{CH})$, are trained from a large corpus by maximum likelihood:

1-char SN:	$P_G(c_i\|LN_{CH})$	=	count(c_i) / count(names)
2-char SN:	$P_G(c_ic_{i+1}\|LN_{CH})$	=	count(c_ic_{i+1}) / count(names)
FN char:	$P_\sigma(c_j\|FN_{CH})$	=	count(c_j) / count(FN chars) of gender σ

We adopted a list of one million personal names in Taiwan to build the probabilistic models. The list contains 476,269 male names and 503,679 female names. There are only 953 surnames and 4,000 more first name characters seen in the name list. For those unseen surnames and first name characters, we assign them a small probability (10^{-1000}, tuned by experiments) to avoid the zero probability problem.

The next step is to build the Chinese name format probability $P_G(\pi \mid \mathrm{S}_{\mathrm{CHname}})$. Since it is about the probability of a name format appearing in an article, the distribution is quite different from the ones observed in the list of one million personal names. A person is often mentioned in an article by his or her title, *e.g*. “Prof. 林” (“Prof. Lin”) or “Mr. 諸葛” (“Mr. Zhu-Ge). When referring to a person in a novel or a letter, it is quite natural to give his or her first name instead of his or her full name. Such cases cannot be captured inside the one million personal names list. Therefore, we need another corpus to train this model.

Personal names in the Academia Sinica Balanced Corpus (*Sinica Corpus* hereafter) are marked as proper nouns (POS-tagged as Nb). We extracted all of the proper nouns in the Sinica Corpus that matched any name format and assumed them to be personal names. These names occur in real documents; thus, they can satisfy our need. The precedence of format matching is defined as follows. Every personal name can only be matched to one format.

1-char word：s > n > not-Chinese-personal-name
2-char word：dd > sn > nn > not-Chinese-personal-name
3-char word：ddn > snn > ssn > not-Chinese-personal-name
4-char word：ddnn > ssnn > not-Chinese-personal-name
5-char word：not-Chinese-personal-name

Nevertheless, for the reason that some common characters are uncommon surnames, it is possible to identify a proper noun of some other type incorrectly as a personal name, such as “中興號” (“Zhong Xing Hao,” a bus company name) where “中” (“Zhong”) is also a surname. In order to increase the precision without sacrificing the recall, we used only frequent surnames and first name characters to do the matching. The sets of frequent characters are the ones that dominate 90% of the probabilities in the name generation model, including 64 surnames (陳,林…程), 467 male first name characters (文,明…瀛), and 293 female first name characters (美,淑…吉), together with all of the 2-syllable surnames.

There are two more formats seen in articles: SN+“姓” or SN+“氏”, which call a person or a family, respectively, by the surname only. We denote them as π = ‘p’. After implementing the matching procedure described above, 39,612 of the 92,314 proper nouns in the Sinica Corpus were extracted as personal names. The Chinese name format probabilities are listed in Table 3. Although there may still be false-alarm personal names in the set, we expect the scale of the corpus is large enough that it can still provide relatively accurate information. The identified personal names in the corpus also can be used to build the bigram models related to the special type S_{CHname}, Chinese personal name.

Table 3. The Chinese name format probabilities

Format Probability	Count	Prob.	Format Probability	Count	Prob.
$P_G(\pi=\text{‘s’}\|S_{CHname})$	5,431	13.71%	$P_G(\pi=\text{‘ddn’}\|S_{CHname})$	126	0.32%
$P_G(\pi=\text{‘n’}\|S_{CHname})$	815	2.06%	$P_G(\pi=\text{‘snn’}\|S_{CHname})$	19,454	49.11%
$P_G(\pi=\text{‘p’}\|S_{CHname})$	487	1.23%	$P_G(\pi=\text{‘ssn’}\|S_{CHname})$	58	0.15%
$P_G(\pi=\text{‘dd’}\|S_{CHname})$	46	0.12%	$P_G(\pi=\text{‘ddnn’}\|S_{CHname})$	24	0.06%
$P_G(\pi=\text{‘sn’}\|S_{CHname})$	2,845	7.18%	$P_G(\pi=\text{‘ssnn’}\|S_{CHname})$	61	0.15%
$P_G(\pi=\text{‘nn’}\|S_{CHname})$	10,265	25.91%	Total	39,612	

An example is given here to illustrate how the probability of a personal name is determined. The word “張德培,” (“Michael Te Pei Chang”) matches two name formats, $\boldsymbol{\pi}$ = {‘snn’, ‘ssn’}, since both “張” (“Chang”) and “德” (“Te”) are possible surnames. Genders options are male and female, *i.e.* $\boldsymbol{\sigma}$ = {M, F}. The most probable one is a male name with the format ‘snn’.

Name: 張德培		
π	σ	Probability
snn	M	$\log(P_G(\text{張}\mid LN_{CH})\times P_M(\text{德}\mid FN_{CH})\times P_M(\text{培}\mid FN_{CH})\times P_G(\pi=\text{‘snn’}\mid S_{CHname}))$ $= (-1.26) + (-1.87) + (-2.74) + (-0.31) = -6.18$
snn	F	$\log(P_G(\text{張}\mid LN_{CH})\times P_F(\text{德}\mid FN_{CH})\times P_F(\text{培}\mid FN_{CH})\times P_G(\pi=\text{‘snn’}\mid S_{CHname}))$ $= (-1.26) + (-2.89) + (-3.27) + (-0.31) = -7.73$
ssn	M	$\log(P_G(\text{張}\mid LN_{CH})\times P_G(\text{德}\mid LN_{CH})\times P_M(\text{培}\mid FN_{CH})\times P_G(\pi=\text{‘ssn’}\mid S_{CHname}))$ $= (-1.26) + (-6.02) + (-2.74) + (-2.82) = -12.84$
ssn	F	$\log(P_G(\text{張}\mid LN_{CH})\times P_G(\text{德}\mid LN_{CH})\times P_F(\text{培}\mid FN_{CH})\times P_G(\pi=\text{‘ssn’}\mid S_{CHname}))$ $= (-1.26) + (-6.02) + (-3.27) + (-2.82) = -13.37$

3.2 Japanese Personal Name Identification

When a Japanese name occurs in an article written in Chinese, there are two ways to write the name. In earlier days, when Traditional Chinese was usually encoded in BIG5, a Japanese name normally was written in its corresponding Traditional Chinese characters, such the name “滝沢秀明,” Hideaki Takizawa, a Japanese performer, would be written as “瀧澤秀明” in Traditional Chinese. Nowadays, many documents are encoded in Unicode, so Japanese Kanji can be directly used in a Traditional Chinese article. Our word segmentation approach wants to identify both cases.

The format of a Japanese personal name is SN+FN, just like a Chinese name. Nevertheless, the length of a Japanese surname varies from one to three Kanji characters, as does the length of the first name part. Sometimes, a name is directly written in Katakana or Hiragana with various lengths. The number of Kanji or Kana characters in a Japanese name is strongly correlated to the number of syllables. Due to the lack of related knowledge, we only deal with the names written in all Kanji and leave the cases of names including Kana as a future work, although Kana can be detected easily by Unicode ranges.

***Table 4. Japanese name formats* (*surnames are underlined*)**

Format	SN	FN	SN+FN
Example	木村 長谷川	理恵 新一	伊藤由奈 高橋留美子

As the length of Japanese names varies considerably, we only adopt three name formats, SN-only, FN-only, and SN+FN, without regarding the number of characters inside the first name part, as listed in Table 4. We know that there is no double surname in Japan.

From the experience of Chinese name processing, we know that a list of Japanese surnames and a large collection of Japanese personal names are needed in order to build name generation probability models. Also, we have to find a corpus of Chinese articles containing

Japanese names in order to build the format probability model as well as the special-type bigram probability. The probability of a Japanese personal name is defined as follows.

$$P_G(w \mid S_{JPname}) = \max_{\pi} P_G(w \mid \pi) P_G(\pi \mid S_{JPname}) \tag{5}$$

The notations in Eq. 5 are defined as the same as in Eq. 4. One difference is that, because we do not have a large training corpus for different genders, the factor of gender in the name generation probability is omitted. Table 5 lists the definitions of each probability, where *m* and *n* are integers between 1 and 3, 'S' denotes the surname part, and 'F' denotes the first name part. Surnames and first names are also assumed to be independent, as are the characters inside a first name part.

Table 5. Definitions of the Japanese name probabilities for every format

Format	Name Generation Probability $P(w\|\pi)$	Format Probability
SN	$P_G(c_{1...}c_m\|LN_{JP})$	$P_G(\pi=\text{'S'}\|S_{JPname})$
FN	$P_G(c_1\|FN_{JP})\times\ldots\times P_G(c_n\|FN_{JP})$	$P_G(\pi=\text{'F'}\|S_{JPname})$
SN+FN	$P_G(c_{1...}c_m\|LN_{JP})\times P_G(c_{m+1}\|FN_{JP})\times\ldots\times P_G(c_{m+n}\|FN_{JP})$	$P_G(\pi=\text{'SF'}\|S_{JPname})$

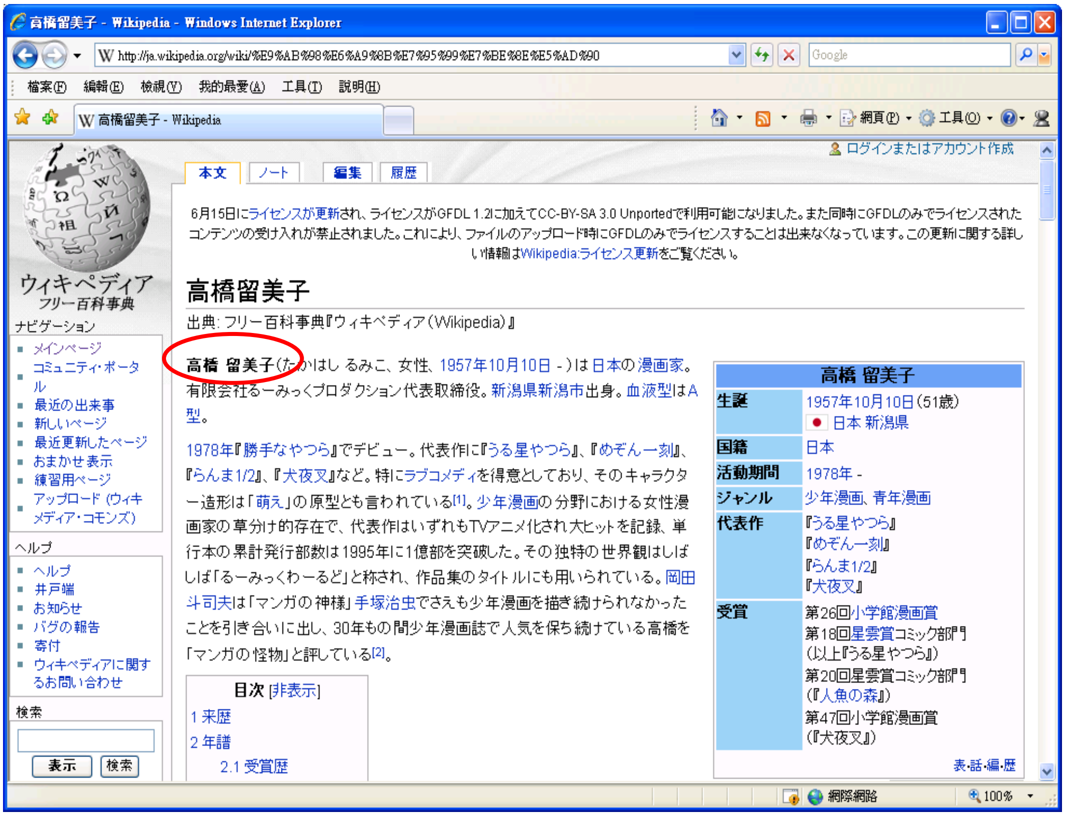
Japanese surnames were collected from a website called "日本の苗字七千傑"[6] (7,000 Surnames in Japan). This website provides 8,603 Japanese surnames along with their populations, where data came from the 117 million costumers of NTT, a Japanese Telecom company. The population data can be used to measure the distributions of the surnames. Nevertheless, according the Wikipedia entry "日文姓名,"[7] there are more than 140 thousand Japanese surnames, far more than we have collected. No complete list is available so far. Moreover, we still need another data set to train the probabilities of first name characters.

All of the Japanese Wikipedia entries that deliver biographies of persons were extracted for learning Japanese personal name distributions. In a Wikipedia page, the title of the entry will also be mentioned again in the text and marked in bold type. The surname part is often separated from the first name part by a space, as in the example of the entry "高橋留美子" ("Rumiko Takahashi"), shown in Figure 1. By detecting such kinds of strings, we can gather many Japanese names in a short time.

Nevertheless, Chinese celebrities may also become entries in the Japanese Wikipedia, such as "王建民" ("Chien-Ming Wang") or "曾國藩" ("Zeng Guofan"). We filtered out the names with a known Chinese surname and a first name part less than three characters. After processing the entire Japanese Wikipedia dumped on Jan 24, 2009 by the methods described above, 65,778 different Japanese names were extracted, including 12,907 surnames and 2,320

[6] http://www.myj7000.jp-biz.net

[7] http://zh.wikipedia.org/wiki/日文姓名


高橋留美子 - Wikipedia - Windows Internet Explorer
http://ja.wikipedia.org/wiki/%E9%AB%98%E6%A9%8B%E7%95%99%E7%BE%8E%E5%AD%90
Google
檔案(F) 編輯(E) 檢視(V) 我的最愛(A) 工具(T) 說明(H)
高橋留美子 - Wikipedia
網頁(P) 工具(O)
ログインまたはアカウント作成
本文 ノート 編集 履歴
ウィキペディア
フリー百科事典
ナビゲーション
メインページ
コミュニティ・ポータル
最近の出来事
新しいページ
最近更新したページ
おまかせ表示
練習用ページ
アップロード (ウィキメディア・コモンズ)
ヘルプ
ヘルプ
井戸端
お知らせ
バグの報告
寄付
ウィキペディアに関するお問い合わせ
検索
表示 検索
網際網路
100%

6月15日にライセンスが更新され、ライセンスがGFDL 1.2に加えてCC-BY-SA 3.0 Unportedで利用可能になりました。また同時にGFDLのみでライセンスされたコンテンツの受け入れが禁止されました。これにより、ファイルのアップロード時にGFDLのみでライセンスすることは出来なくなっています。この更新に関する詳しい情報はWikipedia:ライセンス更新をご覧ください。

高橋留美子

出典: フリー百科事典『ウィキペディア（Wikipedia）』

高橋 留美子（たかはし るみこ、女性、1957年10月10日 - ）は日本の漫画家。有限会社るーみっくプロダクション代表取締役。新潟県新潟市出身。血液型はA型。

1978年『勝手なやつら』でデビュー。代表作に『うる星やつら』、『めぞん一刻』、『らんま1/2』、『犬夜叉』など。特にラブコメディを得意としており、そのキャラクター造形は「萌え」の原型とも言われている[1]。少年漫画の分野における女性漫画家の草分け的存在で、代表作はいずれもTVアニメ化され大ヒットを記録、単行本の累計発行部数は1995年に1億部を突破した。その独特の世界観はしばしば「るーみっくわーるど」と称され、作品集のタイトルにも用いられている。岡田斗司夫は「マンガの神様」手塚治虫でさえも少年漫画を描き続けられなかったことを引き合いに出し、30年もの間少年漫画誌で人気を保ち続けている高橋を「マンガの怪物」と評している[2]。

目次 [非表示]

高橋 留美子	
生誕	1957年10月10日（51歳） ● 日本 新潟県
国籍	日本
活動期間	1978年 -
ジャンル	少年漫画、青年漫画
代表作	『うる星やつら』 『めぞん一刻』 『らんま1/2』 『犬夜叉』
受賞	第26回小学館漫画賞 第18回星雲賞コミック部門 （以上『うる星やつら』） 第20回星雲賞コミック部門 （『人魚の森』） 第47回小学館漫画賞 （『犬夜叉』）

表·話·編·歴

Figure 1. The Wikipedia entry page "*高橋留美子*"

first name Kanji. Table 6 lists the frequencies of these first name Kanji, where the name generation probabilities $P_G(c_j|FN_{JP})$ are listed in the third column and the accumulated probabilities are in the fourth column.

Table 6. Frequencies of the Japanese first name Kanji

FN Kanji	Freq	$P_G(c_j\|FN_{JP})$	Accm Prob.	FN Kanji	Freq	$P_G(c_j\|FN_{JP})$	Accm Prob.
子	4,821	3.60%	3.60%	亨	46	0.03%	89.99%
一	3,358	2.50%	6.10%	瑞	46	0.03%	90.03%
郎	3,237	2.41%	8.52%	…	…	…	…
美	2,230	1.66%	10.18%	褒	1	0.00%	99.99%
正	1,741	1.30%	11.48%	焔	1	0.00%	100.00%
…	…	…	…	Totally 2,320 Kanji; total freq = 134,055			

Many surnames collected from the Japanese Wikipedia did not appear in the surname list of "日本の苗字七千傑". The two lists were merged and resulted in a list of 15,702 surnames.

The population data provided by “日本の苗字七千傑” or the frequencies in Wikipedia were used to estimate the generation probabilities of the surnames, as listed in Table 7. Note that surnames from “佐藤” to “高井良” come from “日本の苗字七千傑,” and the surnames after “斉藤” were collected from Wikipedia.

Table 7. Population of Japanese surnames

SN	Freq	Gen. Prob. $P_G(c_{1...}c_m\|LN_{JP})$	SN	Freq	Gen. Prob. $P_G(c_{1...}c_m\|LN_{JP})$
佐藤	1928000	1.65%	高井良	760	6.49×10^{-6}
鈴木	1707000	1.46%	斉藤	111	9.47×10^{-7}
高橋	1416000	1.21%	三遊亭	106	9.05×10^{-7}
田中	1336000	1.14%	…	…	…
渡辺	1135000	0.97%	城土	1	8.54×10^{-9}
伊藤	1080000	0.92%	駒尾	1	8.54×10^{-9}
…	…	…	Totally 15,702 surnames; total = 117,156,792		

The Japanese name format probability $P_G(\pi \mid S_{JPname})$ was also built by detecting Japanese names in the Sinica Corpus, but only on those proper nouns that were not determined to be Chinese names. Moreover, since the Japanese names in the Sinica Corpus are encoded in Traditional Chinese characters, the matching procedure also includes corresponding Kanji-mapping, which will be explained in Section 4.2.

When extracting Japanese names in the Sinica Corpus, only the 437 first name Kanji (子, 一…瑞), which cover 90% of the probabilities, are used, along with the whole Japanese surname set. The preference of the formats is SN+FN > SN > FN. Each name matched one format at most. After doing so, 4,849 of the 92,314 proper nouns in the Sinica Corpus were extracted as Japanese names. They were used to build the format probability model (as listed in Table 8) as well as the special-type bigram probability for the Japanese name type S_{JPname}. In our experience, however, the format FN-only often suggests too many incorrect candidates and harms the performance of word segmentation. In the end, we elected not to use it.

Table 8. Japanese name format probabilities

Format Probability	$P_G(\pi=\text{‘S’}\|S_{JPname})$	$P_G(\pi=\text{‘F’}\|S_{JPname})$	$P_G(\pi=\text{‘SF’}\|S_{JPname})$	Total
Frequency	718	1,120	3,011	4,849
Probability	14.90%	23.24%	62.48%	

An example is given here to illustrate how the probability of a personal name is determined. The name “滝沢光” matches the Japanese name format in two ways: “滝沢” (“Takizawa”) as a surname and “光” (“Hikaru”) as a first name, or “滝” (“Taki”) the surname and “沢光”

(“Sawahikari”[8]) the first name. The highest probability suggests “滝沢” as a surname and “光” as a first name.

Name: 滝沢光	
Format	Probability
SN+FN	$\log (P_G(\text{滝沢} \mid LN_{JP}) \times P_G(\text{光} \mid FN_{JP}) \times P_G(\pi=\text{‘SN’} \mid S_{JPname}))$ = (-7.35) + (-5.15) + (-0.076) = -12.576
SN+FN	$\log (P_G(\text{滝} \mid LN_{JP}) \times P_G(\text{沢} \mid FN_{JP}) \times P_G(\text{光} \mid FN_{JP}) \times P_G(\pi=\text{‘SN’} \mid S_{JPname}))$ = (-10.70) + (-9.40) + (-5.15) + (-0.076) = -25.326

4. Character Variant Handling

This section discusses three cases where character variants may be used: (1) a Japanese name written in its corresponding Chinese characters (*e.g*. “滝沢秀明” vs. “瀧澤秀明,” Hideaki Takizawa); (2) equivalent words in variant forms (*e.g*. “裡面” vs. “裏面,” inside); (3) Simplified Chinese terms (*e.g*. “體育館” vs. “体育馆”, the gym) appearing in a Traditional Chinese article. Although the last two cases are not often seen, especially the third case (which could not happen until Unicode was invented), we still propose approaches to handle these cases at the same time for the possibility of building a multilingual environment.

4.1 Mapping of Character Variants

A mapping table between the character variants is required for handling the three cases introduced in the previous paragraph. For Japanese names, we need a list of Japanese Kanji and their corresponding Chinese characters. For word variants, a list of the equivalent Chinese character set is necessary. The mapping between Simplified Chinese terms and the corresponding Traditional Chinese ones requires mapping between the two character sets, which is more easily acquired because there are many kinds of software providing such a mapping function.

We do not know of any well-known Japanese-Chinese Kanji mapping tables. To construct one, we adopted the character variant list[9] developed by Prof. Koichi Yasuoka and Motoko Yasuoka in the Institute for Research in Humanities, Kyoto University. There are 8,196 character variant pairs collected in the list. Following the equivalent relationship, we grouped characters in the list into many character-variant clusters. Some examples of character-variant clusters are given here.

[8] In fact, “沢光” (“Sawahikari”) is a Japanese surname and rarely used as a first name.

[9] http://kanji.zinbun.kyoto-u.ac.jp/~yasuoka/ftp/CJKtable/UniVariants.Z

丰 豊 豐 靊 靊
秇 蓺 蕓 藝
乹 乾 乾 干 漧

Note that these variants are equivalent only in some cases. Take the first cluster illustrated above as an example. The character “豊” is Japanese Kanji and “丰” is a Simplified Chinese character, and they both correspond to the Traditional Chinese character “豐”. Nevertheless, “豊” (ritual vessel) and “丰” (elegance) are also legal Traditional Chinese characters that have different meanings from the one of “豐” (prosperous).

In each character-variant cluster, one Traditional Chinese character (if any) is chosen to be the *corresponding* character. If there is more than one Traditional Chinese character in a cluster, the most frequent one is chosen. The frequencies of characters are provided by the Table of Frequencies of Characters in Frequent Words[10] (常用語詞調查報告書之字頻總表) published by the Taiwan Ministry of Education in 1998. Again, considering the first cluster in the examples above, the three characters “丰,” “豊,” and “豐” are all Traditional Chinese characters. “豐” is the most frequent one; hence, it is chosen as the corresponding character of this cluster. By doing so, not only do the Japanese Kanji “豊” and the Simplified Chinese character “丰” have a corresponding Traditional Chinese character, but also the infrequent variants “靊” and “靊” can have a frequent corresponding character.

There are many issues in variant mapping. First, the Traditional Chinese set is larger than the BIG5 character set. Relatively infrequent Traditional Chinese characters, such as “靊,” are not seen in the BIG5 set. Since we are looking for the most frequent Traditional Chinese character, this will not become a problem.

Another issue is the time when two variant characters can be regarded as equivalent. As we have mentioned, the character “豊” is equivalent to “豐” only when it is used as Japanese Kanji. Its meaning in Traditional Chinese is a ritual vessel in ancient times (*cf*. Revised Mandarin Chinese Dictionary[11], 重編國語辭典修訂本), which is completely different from the current meaning of “豐” (prosperous). This would be an interesting future topic.

4.2 Finding Corresponding Chinese Characters for Japanese Kanji

When extracting Japanese personal names inside the Sinica Corpus (as described in Section 3.2), the mapping between Japanese Kanji and Traditional Chinese characters is necessary. Characters in the tables of Japanese surnames and first name Kanji need to be transformed into Traditional Chinese first.

[10] http://www.edu.tw/files/site_content/M0001/87news/index.htm

[11] http://dict.revised.moe.edu.tw/cgi-bin/newDict/dict.sh?cond=%E0T&pieceLen=50&fld=1&cat=&ukey=1838907571&op=&imgFont=1

Each Kanji in a Japanese surname was changed into its corresponding Traditional Chinese character found by the method explained in Section 4.1. For example, the surname “滝沢” (Takizawa) was changed into “瀧澤” and “中曽根” (Nakasone) was changed into “中曾根”. The newly created surnames were merged into the original Japanese surname table, and they shared the same probabilities with the original Japanese surnames. If at least one Kanji character in a surname did not have a corresponding Traditional Chinese character (*e.g.* “畑” in the surname “古畑,” Huruhata), no new surname would be created. The first name Kanji table was expanded in a similar way, along with the assignment of the probabilities.

Merging a newly created term into the name probability table makes our system capable of identifying various methods of name writing at the same time. Our system can identify the two equivalent names in the sentence “滝沢聡就是瀧澤聰” (which means, “滝沢聡 then is 瀧澤聰”). We can see that “滝沢” and “瀧澤” can be found in the Japanese surname table, just as “聡” and “聰” are found in the Japanese first name table. Both “滝沢聡” and “瀧澤聰” are proposed as word candidates that are Japanese names and share the same probability.

Following the same idea, if we further expand the correspondent relationship to the Simplified Chinese character set, it will be possible to understand the sentence “滝沢聡和泷泽聪都是瀧澤聰” (“滝沢聡 and 泷泽聪 both are 瀧澤聰”), where “滝沢聡” is in Japanese, “滝沢聡” is in Simplified Chinese, and “瀧澤聰” is in Traditional Chinese. This part has not yet been implemented but is quite promising.

4.3 Generating Word Variants

In order to identify word variants written either in character variants or in Simplified Chinese, we expanded the dictionary vocabulary by changing the characters in a Traditional Chinese word into their character variants (including Simplified Chinese characters). For example, given a Traditional Chinese word, ABC, each character is searched in the character-variant clusters introduced in Section 4.1. Every character variant found in the character-variant clusters is used to enumerate all possible word variants. Supposing that A’, A”, B’, and C’ are variants of the characters A, B, and C, the following word variants will be enumerated: A’BC, AB’C, ABC’, A’B’C, AB’C’, A’BC’, A’B’C’, A”BC, A”B’C, A”BC’, and A”B’C’.

The newly enumerated word shares the same probability as its original form. Instead of merging the word variants and attaining a large dictionary, we assigned each group of the word variants a unique ID and indexed the bigram probability table (for word segmentation) by the group IDs.

Since the mapping between Simplified Chinese characters and Traditional Chinese characters is not one-to-one, there may be identical word variants enumerated from different words. For example, the Simplified Chinese word variants of “白面” (white-faced) vs. “白麵”

(white noodles) are both “白面,” and the Simplified Chinese word variants “改制” (rule changing) and “改製” (producing in a different model) are the same term “改制,” too. To determine the final probability of an ambiguous word variant, we experimented on three strategies where the final probability is the maximum, the minimum, or the sum of all of the probabilities of the original words. Section 5.4 reveals the results of this experiment.

5. Experiment

5.1 Experimental Data and Evaluation Metrics

The experimental data for word segmentation is the Academia Sinica Balanced Corpus, Version 3.0[12]. The Sinica corpus is designed for language analysis purposes. Words in a sentence are separated by spaces and tagged with their POSs. The documents are written in Modern Mandarin and collected from different domains and topics. There are 316 files containing 743,718 sentences.

Our evaluation was done by 5-fold cross-validation. The 316 files were divided into 5 sets. Each set was used as the test set iteratively when the other sets were used as the development set to construct the lexicon and train probability models. The number of sentences in each set is given in Table 9.

Table 9. Number of sentences in the experimental data

File ID	Test Set ID	No of Files	Sentences	Known Words	Unknown Words
000~065	ASBCset0	66	148,575	146,477	15675
066~129	ASBCset1	64	149,713	146,275	15877
130~183	ASBCset2	54	148,870	146,634	15518
184~244	ASBCset3	61	148,012	146,024	16128
245~315	ASBCset4	71	148,548	146,004	16148

The performance of word segmentation was evaluated by the following metrics, precision, recall, F-measure, and BI score:

$$precision = \frac{\text{correct words being segmented}}{\text{number of words segmented by the system}} \tag{6}$$

$$recall = \frac{\text{correct words being segmented}}{\text{number of words segmented in the test set}} \tag{7}$$

$$\text{F-}measure = \frac{2 \times recall \times precision}{recall + precision} \tag{8}$$

[12] http://godel.iis.sinica.edu.tw/CKIP/20corpus.htm

$$\text{BI-score} = \frac{\text{correct BI labels}}{\text{number of total characters in the test set}} \tag{9}$$

The BI-score labels are defined as follows. Given a sentence, each character is labeled as B (at the beginning of a word) or I (inside a word) according to the segmentation in the test set or the segmentation generated by the system. The ratio of correct BI labels also reveals the performance of a word segmentation system.

When evaluating using 5-fold cross-validation, we used micro-averaging to calculate the scores. That is, the values of the denominators and the numerators of precision, recall, and BI-score are the sums over the five experiment sets.

5.2 Word Segmentation Baseline Performance

This section shows the performance of our basic-model word segmentation system. System Sys1a uses only the known-word lexicon with bigram probability model. System Sys1b integrates the special-type word generation rules, including address, date, time, monetary, percentage, fraction, foreign string, and Internet address, as introduced in Section 2.2. The Sys2 systems further integrate the numbers, including Arabic and Chinese numbers. In order to see the impact of directly adopting the boundary of a number candidate, we experimented on two strategies for Sys2, denoted as Sys2a and Sys2b. As shown in Table 10, Sys1b performs better because of the integration of special-type word generation rules. The maximum-likelihood probability model for numbers is also a better choice.

- **Sys2a: Number generation probability is set to be 1**
- **Sys2b: Number generation probability is trained by maximum likelihood**

Table 10. Performance of the basic word segmentation integrated with special-type word generation rules

System	**R**	**P**	**F**	**BI**
Sys1a	95.66	92.72	94.16	96.96
Sys1b	95.87	93.31	94.57	97.20
Sys2a	95.97	93.57	94.76	97.30
Sys2b	**96.16**	**93.68**	**94.90**	**97.38**

5.3 Experiments on Handling Chinese and Japanese Personal Names

After integrating the Chinese personal name generation rules, the special-type probability for Chinese names is also employed. The difference between our work and Chen *et al.* (1998) is the use of Chinese name format probability and allowing personal names without surnames. Three systems were designed to observe the impact.

- **Sys3a: Using the Chinese name special-type probability, but not the format π = 'nn' and the format probability**
- **Sys3b: Using the Chinese name special-type probability with the format π = 'nn' but not the format probability**
- **Sys3c: Using the Chinese name special-type probability with the format π = 'nn' and the format probability**

All Sys3 systems are based on Sys2b. The evaluation results are shown in Table 11. We can see that all of these methods (using the special-type probability for Chinese name, the name format of FN-only, and the Chinese name format probability) improve the performance. This confirms the success of name formats in personal name recognition and word segmentation.

Table 11. Performance after integrating Chinese name processing

System	**R**	**P**	**F**	**BI**
Sys3a	96.39	94.97	95.68	97.90
Sys3b	96.42	95.49	95.95	98.05
Sys3c	**96.57**	**95.53**	**96.04**	**98.10**

Two systems were designed to observe the effectiveness of the Japanese name special-type probability and the format probability. As the test set is encoded in BIG5, the Japanese name processing is performed under the BIG5 Traditional Chinese character set. Both Sys4 systems are based on Sys3c.

- **Sys4a: Using the Japanese name special-type probability without the format probability**
- **Sys4b: Using both the Japanese name special-type probability and the format probability**

Table 12. Performance after integrating Japanese name processing

System	**R**	**P**	**F**	**BI**
Sys3c	**96.57**	95.53	96.04	98.10
Sys4a	96.54	95.54	96.04	98.10
Sys4b	96.56	**95.56**	**96.06**	98.10

Table 12 illustrates the performance after integrating Japanese name processing. We found that using only the Japanese name special-type probability resulted in a decline of the word segmentation performance, while using both probability models outperformed Sys3c, but not significantly. The reason may be the small number of Japanese names appearing in the Sinica Corpus, as we know that only 4,849 words in the 743,718 sentences were considered to be Japanese names (*cf.* Section 3.2). The improvement of Japanese name processing did not affect the performance of word segmentation significantly.

In order to observe the real performance of Japanese name processing, we designed another experiment. A collection of 109 news articles was prepared, and the Japanese names in it were manually annotated. 862 occurrences of 216 distinct Japanese names were found.

Two kinds of observations were performed. The first one was to verify the ratio of Japanese names being correctly segmented before and after the integration of Japanese name processing. The results are shown in Table 13, which were obtained by applying Sys3c and Sys4b on the 109 documents. This confirms that integrating Japanese name processing greatly improves the success rate.

Table 13. Ratio of Japanese names successfully being segmented

System	**Number of Successfully Segmented Japanese Names**	**Ratio**
Sys3c	154	17.87%
Sys4b	717	83.18%
Total	862	

The second observation is to measure the precision and recall of Japanese name recognition. That is, the ratio of correct ones among the Japanese name candidates proposed by the system (precision) and the ratio of correctly proposed ones among the Japanese names in the test set (recall). The results are listed in Table 14, where both recall and precision are about 75%, which is fair correctness but still needing improvement. This also shows that Japanese name processing is not an easy problem.

Table 14. Precision and recall of Japanese name recognition

System	P	R
Sys4b	74.31% (648/872)	75.17% (648/862)

Some examples of correct and incorrect word segmentation results before and after integrating the Japanese name processing are given here.

Successful examples:

Sys3c	Sys4b	Sys3c	Sys4b
小 林恭二	小林恭二	大 前 研一	大前研一
石原愼 太郎	石原愼太郎	藥師 丸 博子	藥師丸博子

Incorrect examples:

Sys3c	Sys4b	Sys3c	Sys4b
麻布 和 木材	麻布和 木材	瓦斯井 原有	瓦斯 井原有
國小 林佩萱 老師	國 小林 佩萱 老師	廣島 亞運 時	廣島亞運時

5.4 Word Variant Experiments

This section discusses the performance of handling word variants. Unfortunately, we cannot find a suitable test set that contains annotations of character variants. The documents in the Sinica Corpus are encoded in BIG5, a subset of Traditional Chinese characters. There are only a few character variants appearing in the Sinica Corpus.

Two experimental datasets were constructed for the evaluation. The first dataset was a copy of the Sinica Corpus with every character transformed into its Simplified Chinese form (the mapping is unambiguous and can be done by a lot of software). This dataset can be used to verify the ability of Simplified Chinese word handling of a word segmentation system. It can also be used to decide the probabilistic model for homographic variants from different words. The second one was a real corpus written in Simplified Chinese.

As mentioned in Section 4.3, the character mapping from Simplified Chinese to Traditional Chinese is many-to-one. It is possible that a Simplified Chinese word is related to two or more different Traditional Chinese words. Three systems were designed to determine the unigram or bigram probability for such homographic word variants: Sys5a chose the maximum probability among the corresponding Traditional Chinese words, Sys5b chose the minimum, and Sys5c used the sum of the probabilities. Note that Chinese and Japanese name processing also suffers from this problem if the names are written in Simplified Chinese characters. To focus on word variant handling, the experiments were performed without personal name processing. All Sys5 systems were developed based on Sys2b, a system that has not integrated the name processing module. The evaluation results are listed in Table 15. We can see that the method of probability determination does not affect the performance as much, which also shows that the system is capable of dealing with Simplified words in Traditional Chinese text. We chose Sys5a, the one with the maximum values, as our final system.

- **Sys5a: Using the maximal probability of the corresponding source words**
- **Sys5b: Using the minimal probability of the corresponding source words**
- **Sys5c: Using the sum of the probabilities of the corresponding source words**

Table 15. Probability model determination for homographic variants

System	**R**	**P**	**F**	**BI**
Sys5a	96.11	93.53	94.80	97.33
Sys5b	95.95	93.16	94.54	97.21
Sys5c	96.11	93.53	94.80	97.33

The second experiment was done on GHAN 1st Peking University Test Set, a Simplified Chinese word segmentation benchmark. The test set contained 380 sentences. We did not use its development set and lexicon to train our system. Instead, we used Sys5a and the lexicon constructed from the Sinica Corpus. The experimental results show that the performance is worse, where precision is 86.56%, recall is 81.47%, and F-measure is 83.94%. This is because the documents in the Peking University Test Set came from Mainland China, where the vocabulary is quite different from the one in Taiwan. The lower performance is not surprising. The main purpose of this experiment is to show that our system can do word segmentation on documents written in Simplified Chinese with a certain correctness level.

6. Conclusion

In this paper, we propose methods to find word candidates that are Japanese personal names (written in either Japanese Kanji or their equivalent Traditional Chinese characters) or word variants when doing word segmentation. Documents are encoded in UTF-8 so that characters in different languages can appear in the same document. Our word segmentation is based on a bigram probabilistic model, and it integrates the generation rules and probability models for different kinds of special types of words.

When handling Chinese and Japanese personal names, we propose the idea of the name format probability model and discuss how the model can be built. We also propose a method to find corresponding Traditional Chinese characters for Japanese Kanji so that a Japanese name can be detected whenever it is written in a different language. The experimental results show that the name format probability model does improve the performance, and the mappings between Japanese Kanji and Traditional Chinese characters do help to detect Japanese names more successfully.

The size of the Japanese surname list in our system, which contains only 15,702 surnames, is far less than the amount of 140 thousand mentioned in Wikipedia. Nevertheless, once a larger Japanese surname list can be found, it can be easily integrated into our system as long as we assign a small probability to those unseen surnames for smoothing. Furthermore, our knowledge in Japanese name processing is still not sufficient. As a future work, a syllable probabilistic model regarding the pronunciation of a name will be studied. The most important of all is to find a large collection of Japanese names for training.

Using the character variant clusters, Chinese words written in any character variants can be successfully detected as word candidates. Although the set of newly enumerated word variants is large, the computational complexity remains the same if denoting word variants by their group ID and using hash tables to do searching.

Reference

Chen, H.H., Ding, Y.W., Tsai S.C., & Bian, G.W. (1998). Description of the NTU System Used for MET2. In *Proceedings of 7th Message Understanding Conference* (*MUC-7*). Available: http://www.itl.nist.gov/iaui/894.02/related_projects/muc/index.html.

Chien, L.F. (1997). PAT-tree-based keyword extraction for Chinese information retrieval. In *Proceedings of SIGIR97*, 27-31.

Gao, J., Li, M., & Huang, C.N. (2003). Improved Source-Channel Models for Chinese Word Segmentation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics* (*ACL 2003*), 272-279.

羅永聖 (Lo) (2008). 結合多類型字典與條件隨機域之中文斷詞與詞性標記系統研究, Master Thesis, National Taiwan University.

Lu, X. (2007). Combining machine learning with linguistic heuristics for Chinese word segmentation. In *Proceedings of the FLAIRS Conference*, 241-246.

彭載衍 (Peng) and 張俊盛 (Chang) (1993). 中文辭彙歧義之研究－斷詞與詞性標示. In 第六屆中華民國計算語言學研討會論文集 (*ROCLING-6*), 173-194.

Shi, Y. & Wang, M. (2007). A dual-layer CRFs based joint decoding method for cascaded segmentation and labeling tasks. In *Proceedings of International Joint Conference on Artificial Intelligence* (*IJCAI ’07*), 2007, 1707-1712.

Sun, J., Zhou, M., & Gao, J.F. (2003). A Class-based Language Model Approach to Chinese Named Entity Identification. In *International Journal of Computational Linguistics and Chinese Language Processing*, 8(2), 1-28.

Wu, A. & Jiang, Z. (1998). Word segmentation in sentence analysis. In *Proceedings of the 1998 International Conference on Chinese Information Processing*, 169-180.

Zhao, H., Huang, C.N., & Li, M. (2006). An improved chinese word segmentation system with conditional random field. In *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing*, 162-165.