Document Classification Using Domain Specific Kanji Characters Extracted by χ^2 Method

Yasuhiko Watanabe[†] Masaki Murata[‡] Masahito Takeuchi[‡] Makoto Nagao[‡]
[†] Dept. of Electronics and Informatics, Ryukoku University, Seta, Otsu, Shiga, Japan
[‡] Dept. of Electronics and Communication, Kyoto University, Yoshida, Sakyo, Kyoto, Japan
watanabe@rins.ryukoku.ac.jp, {murata, takeuchi, nagao}@pine.kuee.kyoto-u.ac.jp

Abstract

In this paper we describe a method of classifying Japanese text documents using domain specific kanji characters. Text documents are generally classified by significant words (keywords) of the documents. However, it is difficult to extract these significant words from Japanese text, because Japanese texts are written without using blank spaces, such as delimiters, and must be segmented into words. Therefore, instead of words, we used domain specific kanji characters which appear more frequently in one domain than the other. We extracted these domain specific kanji characters by χ^2 method. Then, using these domain specific kanji characters, we classified editorial columns "TENSEI JINGO", editorial articles, and articles in "Scientific American (in Japanese)". The correct recognition scores for them were 47%, 74%, and 85%, respectively.

1 Introduction

Document classification has been widely investigated for assigning domains to documents for text retrieval, or aiding human editors in assigning such domains. Various successful systems have been developed to classify text documents (Blosseville, 1992; Guthrie, 1994; Hamill, 1980; Masand, 1992; Young, 1985).

Conventional way to develop document classification systems can be divided into the following two groups:

1. semantic approach
2. statistical approach

In the semantic approach, document classification is based on words and keywords of a thesaurus. If the thesaurus is constructed well, high score is achieved. But this approach has disadvantages in terms of development and maintenance. On the other hand, in the statistical approach, a human expert classifies a sample set of documents into predefined domains, and the computer learns from these samples how to classify documents into these domains. This approach offers advantages in terms of development and maintenance, but the quality of the results is not good enough in comparison with the semantic approach. In either approach, document classification using words has problems as follows:

1. Words in the documents must be normalized for matching those in the dictionary and the thesaurus. Moreover, in the case of Japanese texts, it is difficult to extract words from them, because they are written without using blank spaces as delimiters and must be segmented into words.
2. A simple word extraction technique generates too many words. In the statistical approach, the dimensions of the training space are too big and the classification process usually fails.

Therefore, the Japanese document classification on *words* needs a high precision Japanese morphological analyzer and a great amount of lexical knowledge. Considering these disadvantages, we propose a new method of document classification on *kanji characters*, on which document classification is performed without a morphological analyzer and lexical knowledge. In our approach, we extracted domain specific kanji characters for document classification by the χ^2 method. The features of documents and domains are represented using the feature space the axes of which are these domain specific kanji characters. Then, we classified Japanese documents into domains by measuring the similarity between new documents and the domains in the feature space.

2 Document Classification on Domain Specific Kanji Characters

2.1 Text Representation by Kanji Characters

In previous researches, texts were represented by significant words, and a word was regarded as a minimum semantic unit. But a word is not a minimum semantic unit, because a word consists of one or more morphemes. Here, we propose the text representation by *morpheme*. We have applied this idea to the Japanese text representation, where a kanji character is a morpheme. Each kanji character has its meaning, and Japanese words (nouns, verbs, adjectives, and so on) usually contain one or more kanji characters which represent the meaning of the words to some extent.

When representing the features of a text by kanji characters, it is important to consider which kanji characters are significant for the text representation and useful for classification. We assumed that these significant kanji characters appear more frequently

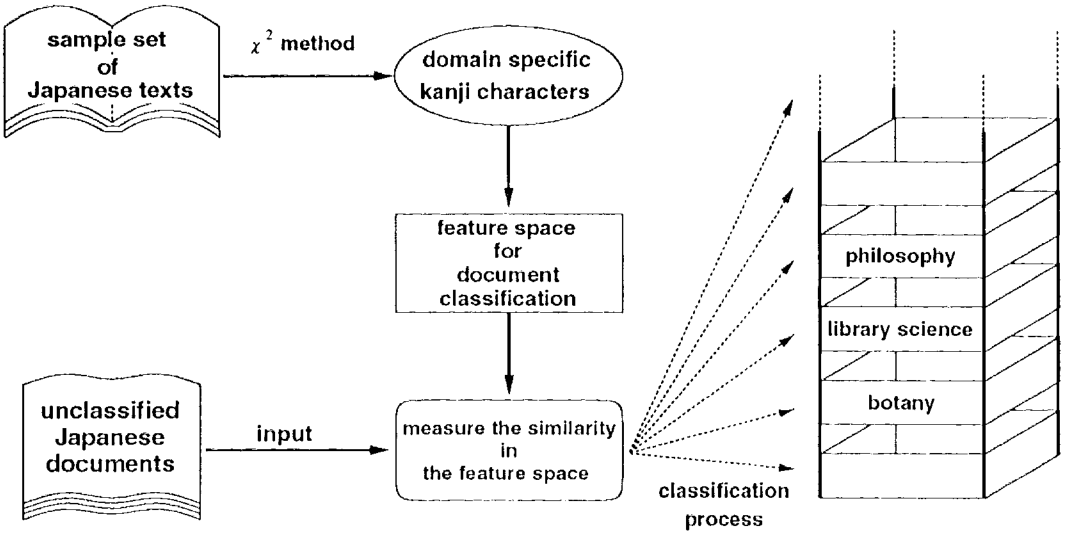


Figure 1: A Procedure for the Document Classification Using Domain Specific Kanji Characters

in one domain than the other, and extracted them by the χ^2 method. From now on, these kanji characters are called the *domain specific kanji characters.*

Then, we represented the content of a Japanese text x as the following vector of domain specific kanji characters:

$$x = (f_1, f_2, \ldots, f_i, \ldots, f_l)\,, \qquad (1)$$

where component f_i is the frequency of domain specific kanji i and l is the number of all the extracted kanji characters by the χ^2 method. In this way, the Japanese text x is expressed as a point in the n-dimensional feature space the axes of which are the domain specific kanji characters. Then, we used this feature space for representing the features of the domains. Namely, the domain v_i is represented using the feature vector of domain specific kanji characters as follows:

$$v_i = (f_1, f_2, \ldots, f_i, \ldots, f_l)\,. \qquad (2)$$

We used this feature space not only for the text representation but also for the document classification. If the document classification is performed on kanji characters, we may avoid the two problems described in Section 1.

1. It is simpler to extract kanji characters than to extract Japanese words.
2. There are about 2,000 kanji characters that are considered necessary for general literacy. So, the maximum number of dimensions of the training space is about 2,000.

Of course, in our approach, the quality of the results may not be as good as in the previous approaches using the words. But it is significant that we can avoid the cost of morphological analysis which is not so perfect.

2.2 Procedure for the Document Classification using Kanji Characters

Our approach is the following:

1. A sample set of Japanese texts is classified by a human expert.
2. Kanji characters which distribute unevenly among text domains are extracted by the χ^2 method.
3. The feature vectors of the domains are obtained by the information on domain specific kanji characters and its frequency of occurrence.
4. The classification system builds a feature vector of a new document, compares it with the feature vectors of each domain, and determines the domain which the document belongs to.

Figure 1 shows a procedure for the document classification using domain specific kanji characters.

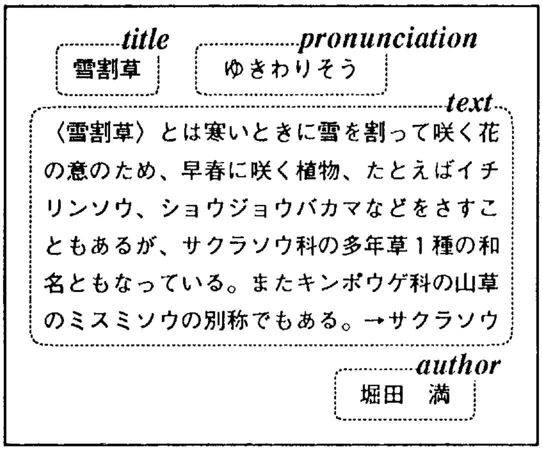
3 Automatic Extraction of Domain Specific Kanji Characters

3.1 The Learning Sample

For extracting domain specific kanji characters and obtaining the feature vectors of each domain, we use articles of "Encyclopedia Heibonsha" as the learning sample. The reason why we use this encyclopedia is that it is published in the electronic form and contains a great number of articles. This encyclopedia was written by 6,722 authors, and contains about 80,000 articles, 6.52×10^7 characters, and 2.52×10^7 kanji characters. An example article of "Encyclopedia Heibonsha" is shown in Figure 2. Unfortunately, the articles are not classified, but there is the author's name at the end of each article and his specialty is notified in the preface. Therefore, we can classify these articles into the authors' specialties automatically.

The specialties used in the encyclopedia are wide, but they are not well balanced [1]. Moreover, some domains of the authors' specialties contain only few

[1] For example, the specialty of Yuriko Takeuchi is Anglo-American literature, on the other hand, that of Koichi Amano is science fiction.


title
pronunciation
text
author

雪割草

ゆきわりそう

〈雪割草〉とは寒いときに雪を割って咲く花の意のため、早春に咲く植物、たとえばイチリンソウ、ショウジョウバカマなどをさすこともあるが、サクラソウ科の多年草１種の和名ともなっている。またキンポウゲ科の山草のミスミソウの別称でもある。→サクラソウ

堀田　満

Figure 2: An Example Article of "Encyclopedia Heibonsha"

articles. So, it is difficult to extract appropriate domain specific kanji characters from the articles which are classified into the authors' specialties.

Therefore, it is important to consider that 206 specialties in the encyclopedia, which represent almost a half of the specialties, are used as the subjects of the domain in the Nippon Decimal Classification (NDC). For example, botany, which is one of the authors' specialties, is also one of the subjects of the domain in the NDC. In addition to this, the NDC has hierarchical domains. For keeping the domains well balanced, we combined the specialties using the hierarchical relationship of the NDC. The procedure for combining the specialties is as follows:

1. We aligned the specialties to the domains in the NDC. 206 specialties corresponded to the domains of the NDC automatically, and the rest was aligned manually.
2. We combined 418 specialties to 59 code domains of the NDC, using its hierarchical relationship. Table 1 shows an example of the hierarchical relationship of the NDC.

However, 59 domains are not well balanced. For example, "physics", "electric engineering", and "German literature" are the code domains of the NDC, and we know these domains are not well balanced by intuition. So, for keeping the domains well balanced, we combined 59 domains to 42 manually.

3.2 Selection of Domain Specific Kanji Characters by the χ^2 Method

Using the value χ^2 of the χ^2 test, we can detect the unevenly distributed kanji characters and extract these kanji characters as domain specific kanji characters. Indeed, it was verified that χ^2 method is useful for extracting keywords instead of kanji characters(Nagao, 1976).

Suppose we denote the frequency of kanji i in the domain j, x_{ij}, and we assume that kanji i is distributed evenly. Then the value χ^2 of kanji i, χ_i^2, is expressed by the equations as follows:

$$\chi_i^2 = \sum_{j=1}^{l} \chi_{ij}^2 \tag{3}$$

$$\chi_{ij}^2 = \frac{(x_{ij} - m_{ij})^2}{m_{ij}} \tag{4}$$

$$m_{ij} = \frac{\sum_{j=1}^{l} x_{ij}}{\sum_{i=1}^{k}\sum_{j=1}^{l} x_{ij}} \times \sum_{i=1}^{k} x_{ij} \tag{5}$$

where k is the number of varieties of the kanji characters and l is the number of the domains. If the value χ_i^2 is relatively big, we consider that the kanji i is distributed unevenly.

There are two considerations about the extraction of the domain specific kanji characters using the χ^2 method. The first is the size of the training samples. If the size of each training sample is different, the ranking of domain specific kanji characters is not equal to the ranking of the value χ^2. The second is that we cannot recognize which domains are represented by the extracted kanji characters using only the value χ^2 of equation (3). In other words, there is no guarantee that we can extract the appropriate domain specific kanji characters from every domain. From this, we have extracted the fixed number of domain specific kanji characters from every domain using the ranking of the value χ_{ij}^2 of equation (4) instead of (3). Not only the value χ_i^2 of equation (3) but the value χ_{ij}^2 of equation (4) become big when the kanji i appears more frequently in the domain j than in the other. Table 2 shows top 20 domain specific kanji characters of the 42 domains. Further, Appendix shows the meanings of each domain specific kanji character of "library science" domain.

3.3 Feature Space for the Document Classification

In order to measure the closeness between an unclassified document and the 42 domains, we proposed a feature space the axes of which are domain specific kanji characters extracted from the 42 domains. To represent the features of an unclassified document and the 42 domains, we used feature vectors (1) and (2) respectively. To find out the closest domain, we measured an angle between the unclassified document and the 42 domains in the feature space. If we are given a new document the feature vector of which is x, the classification system can compute the angle θ with each vector v_i which represents the domain i

$$\theta\,(v_i, x) = \cos^{-1}\left(\frac{v_i \cdot x}{|v_i||x|}\right)$$

and find v_i with

$$\min_i \theta\,(v_i, x)\,.$$

Using this procedure, every document is classified into the closest domain.

Table 1: Division of the Nippon Decimal Classification

5(00)	technology/engineering	class	
54(0)	electrical engineering	code	
548	information engineering	item	
548.2	computers	detailed item	small items
548.23	memory unit	more detailed item	

NDC is the most popular library classification in Japan and it has the hierarchical domains. NDC has the 10 classes. Each class is further divided into 10 codes. Each code is devided into 10 items, which in turn have details using one or two digits. Each domain is assigned by decimal codes.

Table 2: Top 20 Domain Specific Kanji Characters of the 42 Domains

Domain	Domain Specific Kanji Characters (BIG ⟵ the value χ^2_{ij} of equation (4) ⟶ SMALL)																			
library science	書	版	館	冊	庫	本	紙	丁	図	糊	刊	刷	印	巻	帖	誌	文	蔵	折	獄
philosophy	哲	論	学	思	而	想	朱	理	儒	熹	教	的	神	倫	義	派	念	識	孟	彼
psychology	心	我	理	精	習	療	学	眠	能	的	識	児	欲	知	析	神	己	意	験	象
science of religion	寺	教	宗	仏	神	僧	禅	聖	仰	祭	蓮	派	願	信	会	皇	祈	檀	浄	陀
sociology	族	社	会	婚	人	民	猟	畜	儀	住	耕	牧	狩	礼	的	婦	団	呪	姻	男
politics	政	党	治	国	議	権	民	閣	会	主	制	挙	員	選	義	票	軍	争	戦	委
economics	資	労	税	業	済	険	働	金	産	貨	企	価	債	銀	券	経	株	額	財	需
law	法	訴	権	訟	裁	判	条	刑	債	犯	罪	為	憲	務	審	事	罰	項	責	請
military science	艦	戦	軍	隊	砲	撃	兵	弾	攻	爆	銃	核	航	略	敵	艇	空	闘	搭	潜
pedagogy	育	校	教	学	科	児	習	童	制	等	課	師	盲	年	授	稚	塾	設	員	私
commerce	売	商	品	販	卸	店	業	旅	舗	買	購	費	泊	消	宿	取	者	顧	産	企
folklore	餅	漬	煮	染	茶	踊	香	飯	粥	神	煎	醤	俗	菓	豆	料	忌	食	祭	菜
scientific history	学	医	究	研	科	授	術	剖	論	技	博	病	痘	賞	年	蘭	理	物	験	然
mathematics	数	式	幾	関	線	値	算	点	角	何	積	分	微	定	図	次	列	直	限	程
information science	械	憶	処	理	索	計	題	問	算	言	御	識	機	探	報	適	記	解	最	論
astronomy	星	陽	惑	太	天	測	暦	恒	球	月	銀	軌	光	河	鏡	遠	矮	波	宙	午
physics	電	磁	子	波	粒	光	振	力	量	核	度	体	速	熱	荷	射	質	動	気	温
chemistry	酸	溶	塩	素	化	水	液	硫	晶	応	反	合	炭	子	沸	硝	電	錯	結	媒
earth science	岩	気	震	鉱	海	雲	堆	氷	層	噴	候	地	火	石	褶	晶	風	測	雨	温
archeology	墓	墳	器	石	跡	斧	土	棺	葬	遺	銅	鏃	掘	塚	椰	甕	址	陶	陵	竪
biology	酵	胞	遺	酸	細	猿	糖	質	物	生	類	洞	鎖	謝	伝	素	脂	食	雄	体
botany	花	葉	茎	植	咲	枝	萼	草	種	栽	苞	緑	裂	色	培	藻	状	胞	腋	芽
zoology	虫	卵	翅	類	雌	巣	腹	昆	色	褐	魚	肢	吻	幼	殖	雄	鳥	骨	尾	哺
medical science	症	血	病	腫	患	臓	瘍	療	炎	痛	疾	筋	脳	肺	痛	髄	腸	膜	診	骨
engineering	電	船	坑	車	圧	水	機	翼	炉	送	用	速	力	削	路	航	燃	回	舵	波
agriculture	壌	林	培	栽	農	飼	肥	繭	漁	種	畜	苗	森	穫	樹	産	菌	刈	圃	耕
management	郵	送	開	企	営	利	告	業	鉄	便	誌	新	社	益	資	駅	車	輸	報	簿
chemical industry	鋼	溶	料	耐	繊	炉	酸	油	紡	剤	熱	紙	材	維	鋳	脂	金	染	燃	用
machinery	服	衣	袖	着	襟	帽	巾	綿	衿	濯	裾	縫	用	装	袋	丈	毛	襲	繍	飾
architecture	建	築	堂	棟	殿	屋	壁	塔	葺	柱	造	廊	室	住	瓦	居	窓	宅	梁	寺
art	絵	画	彫	美	像	術	描	漆	詩	刻	仏	陶	彩	釉	屏	窯	飾	曼	茶	様
environment	宅	市	都	住	災	街	火	公	計	防	地	画	消	整	設	区	園	施	策	土
printing	写	刷	真	印	版	像	撮	字	紙	光	凹	謄	凸	感	色	用	鋳	肖	稿	罫
music/dance	楽	曲	奏	音	歌	舞	弦	拍	演	箏	声	謡	旋	唄	鼓	踊	作	囃	笛	譜
amusement	劇	映	演	撲	舞	伎	督	郎	画	監	競	俳	棋	賭	座	作	優	踊	瑠	戯
linguistics	語	詞	音	言	字	文	韻	彙	声	漢	舌	母	辞	仮	嗜	英	唇	詩	話	書
Western literature	詩	劇	人	文	説	作	彼	恋	評	悲	家	寓	女	抒	批	魔	叙	編	愛	世
Eastern literature	歌	諧	俳	句	蕉	詩	話	撰	芭	謡	文	詠	説	亭	巻	娼	集	狂	郎	江
geography	川	山	地	県	島	町	南	北	岸	郡	西	港	東	市	峠	泉	湖	丘	麓	口
ancient history	王	帝	前	僭	征	隷	記	奴	民	市	軍	政	族	世	彼	朝	盟	神	文	都
Western history	党	政	民	領	国	革	年	王	議	会	邦	権	帝	世	命	農	軍	争	戦	蜂
Eastern history	氏	藩	荘	幕	朝	郡	皇	府	領	城	町	田	揆	武	臣	姓	国	政	官	惣

4 Document Classification Using Domain Specific Kanji Characters

4.1 Experimental Results

For evaluating our approach, we used the following three sets of articles in our experiments:

1. articles in "Scientific American (in Japanese)" (162 articles)
2. editorial columns in Asahi Newspaper "TENSEI JINGO" (about 2,000 articles)
3. editorial articles in Asahi Newspaper (about 3,000 articles)

Because the articles in "Scientific American (in Japanese)" are not classified, we classified them manually. The articles of "TENSEI JINGO" and the editorial articles are classified by editors into a hierarchy of domains which differ from the domains of the NDC. We aligned these domains to the 42 domains described in Section 3.1. Some articles in thereof contain two or more themes, and these articles are classified into two or more domains by editors. For example, the editorial article "Too Many Katakana Words" is classified into three domains. In these cases, we judge that the result of the automatic classification is correct when it corresponds to one of the domains where the document is classified by editors. Figure 3 , Figure 4, and Figure 5 describe the variations of the classification results with respect to the number of domain specific kanji characters.

4.2 Evaluation

In our approach, the maximum correct recognition scores for the editorial articles and the articles in "Scientific American (in Japanese)" are 74 % and 85 %, respectively. Considering that our system uses only the statistical information of kanji characters and deals with a great amount of documents which cover various specialties, our approach achieved a good result in document classification. From this, we believe that our approach is efficient for broadly classifying various subjects of the documents, e.g. news stories. A method for classifying news stories is significant for distributing and retrieving articles in electronic newspaper.

The maximum recognition scores for "TENSEI JINGO" is 47 %. The reasons why the result is far worse than the results of the other are:

1. The style of the documents

   The style of "TENSEI JINGO" is similar to that of an essay or a novel and it is written in colloquial Japanese. In contrast, the style of the editorial articles and "Scientific American (in Japanese)" is similar to that of a thesis. We think the reason why we achieved the good result in the classification of the editorial articles and "Scientific American (in Japanese)" is that many technical terms are used in there and it is likely that the kanji characters which represent the technical terms are domain specific kanji characters in that domain.

2. Two or more themes in one document

   Many articles of "TENSEI JINGO" contain two or more themes. In these articles, it is usual that the introductory part has little relation to the main theme. For example, the article "Splendid Retirement", whose main theme is the Speaker's resignation of the House of Representatives, has an introductory part about the retirement of famous sportsmen. In conclusion, our approach is not effective in classifying these articles.

   However, if we divide these articles into semantic objects, e.g. chapter and section, these semantic objects may be classified in our approach. Table 3 shows the results of classifying full text and each chapter of a book "Artificial Intelligence and Human Being". Because this book is manually classified into the domain

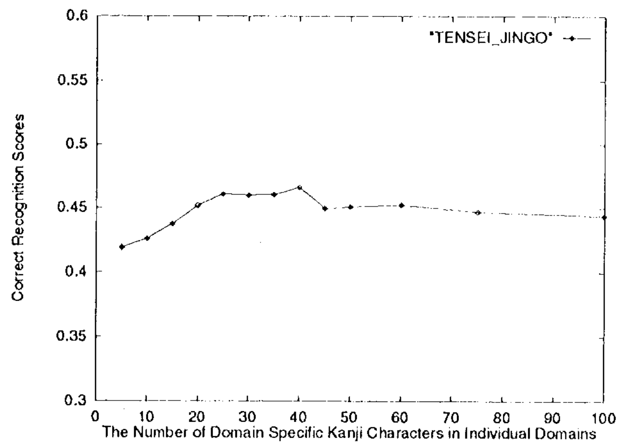


Figure 3: Variations of the Classification Results for "TENSEI JINGO" by the Number of Domain Specific Kanji Characters in Individual Domains

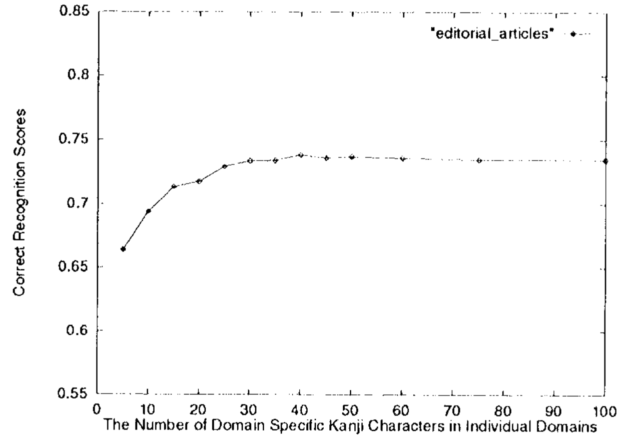


Figure 4: Variations of the Classification Results for the editorial articles by the Number of Domain Specific Kanji Characters in Individual Domains

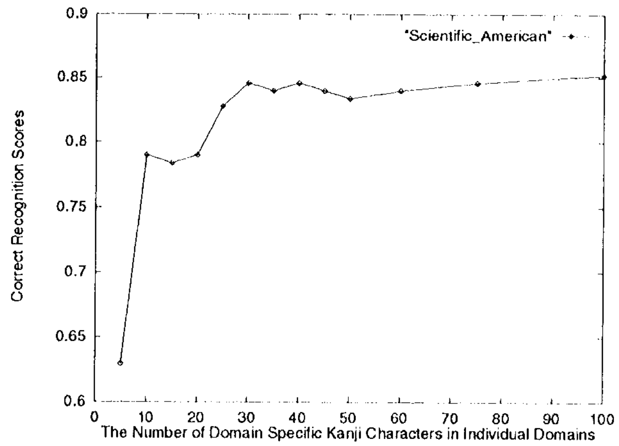


Figure 5: Variations of the Classification Results for "Scientific American (in Japanese)" by the Number of Domain Specific Kanji Characters in Individual Domains

Table 3: A Classification Result of a book "Artificial Intelligence and Human Being"

Chapter	Title	Result
Chapter 1	The Ability of Computers	information science
Chapter 2	Challenge to Human Recognition	information science
Chapter 3	Aspects of Natural Language	linguistics
Chapter 4	What is the Understanding ?	information science
Chapter 5	Artificial Intelligence and Philosophy	psychology
Full Text of "Artificial Intelligence and Human Being"		information science

"information science" in the NDC, it is correct that the system classified this book into the "information science". And it is correct that the system classified Chapter 3 and Chapter 5 into the "linguistics" and "psychology", respectively, because human language is described in Chapter 3 and human psychological aspect is described in Chapter 5.

5 Conclusion

The quality of the experimental results showed that our approach enables document classification with a good accuracy, and suggested the possibility for Japanese documents to be represented on the basis of kanji characters they contain.

6 Future Work

Because the training samples are created without this application in mind, we may be able to improve the performance by increasing the size of the training samples or by using different samples which have the similar styles and contents to the documents. We would also like to study the relation between the quality of the classification result and the size of the documents.

References

Blosseville M.J, Hébrail G., Monteil M.G., Pénot N.: "Automatic Document Classification: Natural Language Processing, Statistical Analysis, and Expert System Techniques used together", SIGIR '92, pp. 51-58, 1992.

Guthrie L., Walker E., Guthrie J.: "DOCUMENT CLASSIFICATION BY MACHINE:Theory and Practice", COLING 94, pp. 1059-1063, 1994.

Hamill K.A., Zamora A.: "The Use of Titles for Automatic Document Classification", Journal of the American Society for Information Science, pp. 396-402, 1980.

Masand B., Linoff G., Waltz D.: "Classifying News Stories using Memory Based Reasoning", SIGIR '92, pp. 59-65, 1992.

Nagao M., Mizutani M, Ikeda H.: "An Automatic Method of the Extraction of Important Words from Japanese Scientific Documents" (in Japanese), Transactions of IPSJ, Vol.17 No.2, pp.110-117, 1976.

Young S.R., Hayes P.J.: "Automatic Classification and Summarization of Banking Telexes", Proceedings of the Second IEEE Conference on AI Applications, pp. 402-408, 1985.

Appendix

The meanings of each domain specific kanji character of the "library science" category are as follows:

書 write; draw; writing, art of writing, calligraphy, penmanship; books, literary work; letter, note

版 printing block, printing plate, wood block; publishing, printing; printing, edition, impression;

館 building, hall, mansion, manor ; suffix of public building (esp. a large building for cultural activities), hall, edifice, pavilion;

冊 counter for books, volumes or copies; bound book, volume, copy;

庫 storehouse, warehouse, storage chamber

本 basis, base, foundation; origin, source, root, beginning; book, volume, work, magazine; this, the same, the present; head, main, principal; real, true, genuine; counter for cylindrical objects (bottles, pencils, etc)

紙 paper; newspaper, periodical, publication

丁 town subsection, city block-size area; counter for dished of food, blocks of tofu, guns; two-page leaf of paper

図 drawing, diagram, plan, figure, illustration, picture; map, chart; systematic plan, scheme, attempt, intention;

糊 paste, glue; starch, sizing

刊 publish; publication, edition, issue

刷 print, put in print; counter for printings

印 (visual sign)seal, stamp, seal impression; sign, mark, symbol, imprint; print; India

巻 volume, book; roll, reel; roll up, roll, scroll, wind, coil

帖 notebook, book, register; counter for quires (of paper), folding screens, volumes of Japanese books, etc.; counter for tatami mats

誌 magazine, periodical, suffix names of magazines of periodicals; write down, chronicle

文 letter, character, script, inscription; writing, composition, sentence, text, document, style; letters, literature, the pen; culture, learning, the arts design; letter, note;

蔵 store, put away, lay by; own, possess, keep (a collection of books); storehouse, storing place, treasury

折 break, be folded, bent; turn (left/right); yield, compromise

獄 prison, jail; hell; lawsuit, litigation