Hindawi Publishing Corporation
Advances in Artificial Intelligence
Volume 2012, Article ID 562604, 18 pages
doi:10.1155/2012/562604

Research Article

A Method for Identifying Japanese Shop and Company Names by Spatiotemporal Cleaning of Eccentrically Located Frequently Appearing Words

Yuki Akiyama[1] and Ryosuke Shibasaki[2]

[1] *Center for Spatial Information Science, The University of Tokyo, Cw-503 Shibasaki Laboratory, 4-6-1 Komaba, Meguro-ku, Tokyo 153-8505, Japan*
[2] *Center for Spatial Information Science, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa City, Chiba 277-8568, Japan*

Correspondence should be addressed to Yuki Akiyama, aki@iis.u-tokyo.ac.jp

Received 23 July 2011; Revised 9 November 2011; Accepted 2 December 2011

Academic Editor: Mohamed Afify

We have developed a method for spatiotemporally integrating databases of shop and company information, such as from a digital telephone directory, spatiotemporally, in order to monitor dynamic urban transformations in a detailed manner. To realize this, an additional method is necessary to verify the identicalness of different instances of Japanese shop and company names that might contain fluctuations of description. In this paper, we discuss a method that utilizes an n-gram model for comparing and identifying Japanese words. The processing accuracy was improved through developing various kinds of libraries for frequently appearing words, and using these libraries to clean shop and company names. In addition, the accuracy was greatly and novelty improved through the detection of those frequently appearing words that appear eccentrically across both space and time. By utilizing natural language processing (NLP), our method incorporates a novel technique for the advanced processing of spatial and temporal data.

1. Introduction

Spatiotemporal changes of shop and company locations have a major effect on the vitality and attraction of urban space. It is a significant challenge to monitor these changes, quantitatively and in as detailed as manner as possible, for use in various fields including urban engineering, geography, and economics. However, it is difficult to comprehensively monitor urban spaces, because much general regional and statistical information (e.g., the population census, commercial statistics) is compiled by separate administrative or city block units.

On the other hand, detailed information on shop and company locations and names can be collected using telephone directories and web information. Fortunately, this is possible in Japan, because of the availability of digital telephone directories and detailed digital maps which can monitor almost all residents and tenants in a given building.

The yearly continuations and changes in tenants or residents can be monitored for a certain location, and we can integrate these data across multiple years. The same can be done for shop and company locations over multiple years, by measuring changes in shop and company names. However, this measure is not easy because of name fluctuations between different two years or different kinds of data. Therefsore, we have been developing a dataset that can monitor the time-series changes of each shop and company and a system that can develop such data as to resolve this challenge [1, 2]. This paper focuses on a particular method of name identification, pertinent to shop and company names—that is, an identification method for Japanese words.

1.1. Previous Studies: About Spatial and Temporal Data Developments. There have been many previous studies that have attempted to monitor changes in urban spaces using time-series information of shops, companies, and buildings.

For example, the locations of open and closed shops were extracted using digital maps and the results of field surveys by Ato et al. [3], and time-series changes of building locations were extracted and applications were developed using digital residential maps by Ai et al. [4]. However, the processing methods used in almost all of these previous studies are inappropriate for large quantities of data (e.g., the whole area of one city or prefecture) because the researchers developed their time-series data by manual, time-intensive processing.

On the other hand, Ito and Magaribuchi have developed a completely automated method of spatiotemporal integration of digital residential maps, which is capable of processing large volumes of data [5]. However, this method has been applied to only one specific kind of digital map. In addition, the method encounters difficulties when dealing with problems of so-called noise word cleaning and local frequently appearing words (to be described below) because of the focus of this study was only Tokyo's 23 wards. As result, we consider that there are some limitations to apply this method over a broad area.

1.2. Previous Studies: About Language Processing. For this study, a method to recognize name entities (i.e., compound noun) is necessary. There have been many previous studies that worked in various ways to develop this method. Florian et al. presented a statistical language-independent framework for identifying and tracking named, nominal, and pronominal references to entities within unrestricted text documents, then chaining them into clusters corresponding to each logical entity present in the text [6]. Tri Tran et al. applied a support-vector-machine- (SVM-) based NER model to the Vietnamese language [7]. Tjong Kim Sang and Meulder processed named entity data from English and German using sixteen different kinds of systems to recognize the entities' identities, and they obtained the best result using a combined learning system that applied Maximum Entropy to each language [8]. In addition, there have been many studies that have attempted to recognize name entities from other kinds of languages [9–13].

There have also been many previous studies focusing on the processing of Japanese words. Sato et al. developed a method to predict the authors of a text based on frequencies of word usage within it [14]. Kawakami and Suzuki presented a method to calculate word similarities in random texts using a decision list [15]. Mishina et al. evaluated word similarities using *n*-grams. Similarly, we will use *n*-grams in this study in order to recognize and identify shop and company names [16].

However, it has been difficult for previous methods to deal with local frequently appearing words (LFAW). Our approach to managing this problem is introduced in Section 3.5 in detail.

1.3. About This Paper. This paper and our system focus on Japanese language processing. Our system can monitor the time-series changes of each shop and company, integrating them to create a dataset containing their names and locations, that is, address, longitude, and latitude across two years spatially and to measure identifications of their names. In this paper, we focus on how to measure identifications of Japanese words.

Table 1: Example of description of Kanji by Hiragana.

	Described by Kanji	Described by Hiragana/Katakana
Japanese characters	日本	にほん
Pronunciations	Nihon	Ni Ho Nn
Meaning in English	Japan	Japan

Table 2: Pronunciations by Chinese and Japanese of the same characters.

Chinese/Japanese character	中	山	本
Pronunciations in Chinese	Zhōng	Shān	Běn
Pronunciations in Japanese (only common readings)	Naka Chūuu	Yama San Sen	Hon Bon Pon Moto

There are two remarkable and novel points our paper introduces. The first is that it utilizes natural language processing (NLP) for the advanced processing of spatial and temporal data. There are few studies that have processed data using NLP in the field of spatial information science. Some studies in this field have partly utilized NLP [17, 18]; however, there are no studies that have utilized NLP for the processing of spatial and temporal data to the same extent as our study. In Japan, Ito and Magaribuchi [5] have accomplished a similar trial to our study, as detailed in Section 1.1. However, their method has been applied only in central Tokyo. Our study is the first to develop a spatiotemporal dataset for throughout Japan. The second is that our method can deal with LFAW. It is a novel method that recognizes so-called pure shop and company names ("Pure" refers to the elements of a character string which identifies a tenant uniquely.) and detects words that are eccentrically-located spatially and temporally (LEAW) and also cleans shop and company names of LFAW.

In Japan, there are many kinds of data that contain name and location information. One of the largest and most complete datasets for shops and companies all over Japan comprises residential and tenant information from digital residential maps (Zenrin CO., Ltd.; in Japanese, "Zyutaku-Chizu") and digital telephone directories (e.g., "Town Page Database" by NTT Business Information Service, Inc. and "Telepoint Data" by Zenrin CO., Ltd.). Other data are data of each companies and enterprises, for example, the quarterly journal of companies and enterprises of Japan, or shop information on the Web collectable by API services. Therefore, our method can be adapted across various fields of data development.

Our system should be able to process data in various regions and times. Therefore, the test data used in the development of our system was the residential and tenant information in the digital residential maps and telephone

Table 3: Example of Japanese without blanks.

	Text	Translation from English to Japanese
Japanese	東京証券取引所	
English	Tokyo Stock Exchange	Tokyo = 東京 Stock = 証券 Exchange = 取引所

Table 4: Examples of description of loanwords by katakana.

		Example 1	Example 2	Example 3
Words	Loanwords	Notebook (En)	Baumkuchen (De)	Château (Fr)
	Japanese	ノートブック	バームクーヘン	シャトー
Pronunciations	Loanwords	nóutbuk	baʊm'kuːxn̩	ʃætóʊ
	Japanese	nōtobukku	Bāmukūhen	Shatō

Table 5: Examples of noise words in shop and company names.

	Example 1	Example 2
Shop and company names		
Japanese	マクドナルド下北沢店	スターバックスコーヒー六本木ヒルズ店
English	McDonalds Shimokitazawa shop	Starbucks Coffee Roppongi Hills shop
Noise words		
Japanese	下北沢店	六本木 ヒルズ店
English	Shimokitazawa shop	Roppongi Hills shop
Kind of noise words	Station/geographic name	Building name

Table 6: Examples of fluctuations of description between old and new names.

Name (in 2005)	Name (in 2000)	Address
20世紀フォックス映画	20世紀フオツクス映画会社	東京都港区六本木3丁目16-33
55ステーション江戸川橋店	55分DPEステーション江戸川橋店	東京都文京区関口1丁目6-10
747サウンドポート新宿南口本店	サウンドポートIN747新宿南口本店	東京都新宿区新宿3丁目36-16
ABC新宿クッキングスタジオ	ABCクッキングスタジオ新宿店	東京都新宿区西新宿1丁目26-2
Antsケントレーディング	ケントレーディングブレイン株式会社	東京都中央区銀座1丁目14-10
ARAトラベル合宿インフォメーション	ARAトラベルガッシュクインフォメーション	東京都新宿区新宿3丁目12-4
auショップ六本木交差点	IDOプラザ六本木交差店	東京都港区六本木4丁目8-7
BA—RU	BA・RU	東京都文京区千石4丁目38-10
B—マリエ	ビーマリエ	東京都港区白金1丁目29-15
CLAN・PaPa	ビアレストランP	東京都文京区根津2丁目11-8
COM陶芸教室	コム(COM)陶芸教室	東京都新宿区西新宿7丁目6-6
Di・マーレ	ディマーレー	東京都台東区浅草橋3丁目20-18
ELEC英語研修所	エレック英語研修所	東京都千代田区神田錦町3丁目20
ESPギタークラフト・アカデミー東京	ESPギタークラフトアカデミー東京	東京都千代田区神田錦町1丁目8
IDC大塚家具・新宿ショールーム	IDC大塚家具新宿ショールーム	東京都新宿区新宿3丁目33-1
KAZUKIスパゲティ専門店	スパゲティ専門店KAZUKI	東京都千代田区内神田1丁目4-12
KOJI・VANCOUVER・赤坂	KOJI・VANCOUVER赤坂	東京都港区赤坂2丁目14-27
MKミッシェルクラン新宿小田急店	ミッシェルクランエムケー	東京都新宿区西新宿1丁目1-3

directory mentioned above, because these data can cover all of Japan with a homogeneous resolution.

2. Characteristic Features of Japanese

Japanese is a language used throughout Japan. There are about 130 million speakers in the world, mainly in East Asia [19].

One of the main characteristics of Japanese is its use of three kinds of characters: hiragana, katakana, and kanji. Hiragana and katakana are phonograms, while kanji are ideograms. The origin of kanji is Chinese characters, while hiragana and katakana are unique characters originating in Japan. Kanji are mainly used to write nouns, roots of verbs and adjectives, and personal names of Japanese and Chinese people. Kanji can be described by hiragana and katakana because the pronunciations of kanji can be written phonetically, as seen in Table 1. In addition, there are multiple pronunciations in kanji, unlike Chinese characters (Table 2). In such cases, the pronunciation of a given kanji is decided by the context of the surrounding words and texts.

A notable characteristic of written Japanese is that it does not have blanks between single words. Because of this, it is difficult to divide one text into its component single words without an adequate understanding of word meaning or class (Table 3). This is a common characteristic of major languages in East Asia. Klein et al. have also pointed this out and tried to recognize Chinese name entities using the character sequences [20].

In addition, one of the interesting features of Japanese is that it typically writes loanwords with similarly pronounced katakana (Table 4). Chinese also has this same feature. When it is difficult to write loanwords with katakana because of inadequate or incompatible phonemic inventories, loanwords are sometimes written in the original languages and script [21].

2.1. Difficulty in Verifying the Identity of Shop and Company Names. It is not easy to verify that two Japanese words are identical or even similar because of the above features of Japanese. For example, character string lengths tend to be longer than English or French, because Japanese is written without blanks between single words. In addition, there are many fluctuations of description, because Japanese uses three kinds of characters and changes word order frequently.

Moreover, one kind of character string that appears frequently in shop and company names is branch names. These strings become noise words, making name identification difficult if a shop and company name contains long geographic or building names (Table 5). We realize the necessity of solving these kinds of problems if we wish to verify the identicalness of different instances of shop and company names adequately.

3. Development

We identify and verify the time-series changes of shops and companies between two different years based on location (i.e., address, longitude, and latitude) and name information (i.e., shop, company, or building names). Then, we can assess the kind of time-series change—that is, continuation, change, emergence, and demise—of each shop or company between different two years, for monitoring purposes.

3.1. Input Data. The input data consisted of name and location information separated by commas (e.g., in csv or txt format) containing an address at minimum. When more specific information is provided in the source data—building names, floors, and room numbers—our system can integrate input data more accurately than without. Figure 1 shows an image of some sample input data and their resultant output.

3.2. Processing Flow. Figure 2 shows the processing flow of our method and all potential results of time-series integration. At first, new and old data are integrated spatially for each shop and company unit. Shops and companies found at the same location are integrated into a set, and after subsequent time-series integration, they are labeled either "continuation" or "change." The time-series results of shop and companies that exist only as new data are labeled "emergence" and those that exist only as old data are labeled "demise."

In this paper, we introduce a method to verify whether or not two spatially integrated names refer to the same tenant, and then to decide whether the time-series change is best classified as "continuation" or "change." The details of spatial integration have been described in our previous studies [1, 2].

In this study, the time-series changes of each shop and company were monitored based on the "name" changes within the same location. In other words, our method monitors time-series changes of buildings and their floors and rooms. Therefore, our method will not identify transfers of shop or company ownership, whether by merger or acquisition. However, we expect that interested parties will be able to track changes of ownership at the same "continuation" location by using other data or statistics more relevant to company mergers and acquisitions, such as the Japan Company Handbook (Toyo Keizai Inc.).

3.3. Verification of Name Identification. It is not easy to verify that a new name and old name refer to the same company at a given location, because simply determining whether each name is exactly the same returns inadequate results. There are subtle fluctuations of description between the names in new data versus old data, even though they may actually be the same shops or companies. Table 6 shows some examples of fluctuations of description between old and new names of the same business. The shops and companies listed in Table 6 were taken from the 2000 and 2005 Tokyo telephone directories. Each shop or company is located at the same location in 2000 and 2005, and this fact can be verified via human manual processing. However, each name is subtly different.

In order to solve this problem, we must meaningfully quantify similarities between the words of the shop and company names.

Table 7: Appearance frequencies of FAW in tenant names in 2005.

FAW in Japanese	FAW in English	Appearance frequency	Ratio of appearance (%)
（株）	Co.	532007	13.88
（有）	Ltd.	322517	8.42
株式会社	Corporation	151421	3.95
有限会社	Limited company	69205	1.81
センター	Center	54179	1.41
（事）	Office	39617	1.03
美容室	Beauty salon	32534	0.85
（営）	Business office	28489	0.74
ビル	Building	27510	0.72
クリニック	Clinic	19028	0.50
東京	Tokyo	18702	0.49
駐車場	Parking	17703	0.46
ハイツ	Heights	17679	0.46
サービス	Service	17542	0.46
クリーニング	Cleaning	17432	0.45
スナック	Snack bar	14308	0.37
コーポ	Cooperative	13716	0.36

▪RThe ratio of appearance is calculated as the appearance frequency divided by the tenant total.
▪The full library contains 963 words.
Tenants in the 2005 residential maps: 3,141,434
Tenants the 2005 telephone directory: 690,183
Total tenants: 3,831,617.

Table 8: Examples of noise words in shop and company names.

Names (Japanese/English)	Noise words	Kind of noise word
株式会社ゼンリン/ Zenrin Co., Ltd.	株式会社/Co., Ltd.	FAW
サーティワンアイスクリーム麻布店/ 31 ice cream Azabu shop	麻布店/Azabu shop	Geographic name
かつや代々木駅前店/ Katsuya in front of Yoyogi station	新橋駅前店/in front of Yoyogi station	Station name
株式会社ホブソンズジャパン西麻布店/ Hobsons Japan Co. Nishiazabu office	株式会社/Co. 西麻布店/Nishiazabu office	FAW Geographic name
アコム株式会社新橋駅前支店/Acom Co., Ltd in front of Shinbashi station	株式会社/Co., Ltd. 新橋駅前支店/in front of Shinbashi station	FAW Station name
喫茶室ルノアール赤坂見附店/ Coffee shop Renoir Akasaka Mitsuke shop	喫茶室/Coffee shop 赤坂見附店/Akasaka Mitsuke shop	FAW Geographic name

In this study, this word quantification has been realized by the "n-gram." The n-gram is one method of natural language processing that can quantify the degree of similarity between different two words [22]. The method has been attracting attention in fields as diverse as literature, linguistics, and computer science [23–25].

We use the bigram (2-gram) to calculate name similarity in this study. The bigram extracts string blocks constructed of 2 characters from new and old names and then compares them. This method can resolve the problem of fluctuations of description. Figure 3 depicts the bigram calculation method as applied to our word similarity problem. A name similarity between word i and word j is defined by

$$S_{ij}{}^{(n)} = \frac{n_{ij}{}^{(n)} + n_{ji}{}^{(n)}}{m_i{}^{(n)} + m_j{}^{(n)}}, \quad (1)$$

where $S_{ij}{}^{(n)}$ is the name similarity between word i and word j, $m_i{}^{(n)}$ is the number of string blocks extracted from word i, $m_j{}^{(n)}$ is the number of string blocks extracted from word j, and $n_{ij}{}^{(n)}$ and $n_{ji}{}^{(n)}$ are the number of string blocks within $m_i{}^{(n)}$ matching within $m_j{}^{(n)}$, and vice versa, respectively.

New data

Name	Location Information			
	Address	Bldg information	Longitude	Latitude
ロッテリア大通店	北海道札幌	大通ビル1階	141.35	43.06
養老の瀧	北海道札幌	大通ビルB1階	141.35	43.06
札幌観光(株)	北海道札幌	大通ビル303号	141.35	43.06
北海道食品	北海道札幌	鉄北会館2A	141.34	43.07
札幌かに道場	北海道札幌	薄野タワー202	141.35	43.05
セイコーマート琴似	北海道札幌	ハイム琴似1F	141.31	43.07

Old data

Name	Location Information			
	Address	Bldg information	Longitude	Latitude
ロッテリア札幌大通店	北海道札幌	大通りビル1F	141.35	43.06
つぼ八大通り店	北海道札幌	大通りビルB1F	141.35	43.06
北洋商事	北海道札幌	大通りビル202	141.35	43.06
北海道食品(株)	北海道札幌	鉄北会館3B号	141.34	43.07
かに道場	北海道札幌	薄野タワー2階	141.35	43.05
セイコーマート	北海道札幌		141.31	43.07

Result of time-series integration

Name in new time	Name in old time	Location information						Result of time-series integration
		Address	Bldg info	floor	Room No.	Lon	Lat	
ロッテリア大通店	ロッテリア札幌大通店	北海道札幌	大通ビル	1		141.35	43.06	Continuation
養老の瀧	つぼ八大通り店	北海道札幌	大通ビル	−1		141.35	43.06	Change
札幌観光(株)		北海道札幌	大通ビル	3	303	141.35	43.06	Emergence
	北洋商事	北海道札幌	大通りビル	2	202	141.35	43.06	Demise
北海道食品		北海道札幌	鉄北会館	2	A	141.34	43.07	Emergence/imigration
	北海道食品(株)	北海道札幌	鉄北会館	3	B	141.34	43.07	from other floor
札幌かに道場	かに道場	北海道札幌	薄野タワー	2	202	141.35	43.05	Continuation (add floor info)
セイコーマート琴似	セイコーマート	北海道札幌	ハイム琴似	1		141.31	43.07	Continuation (add bldg info)

Note: All information in this table is fictional.

Figure 1: Image of sample input data and resultant output (time-series integration).

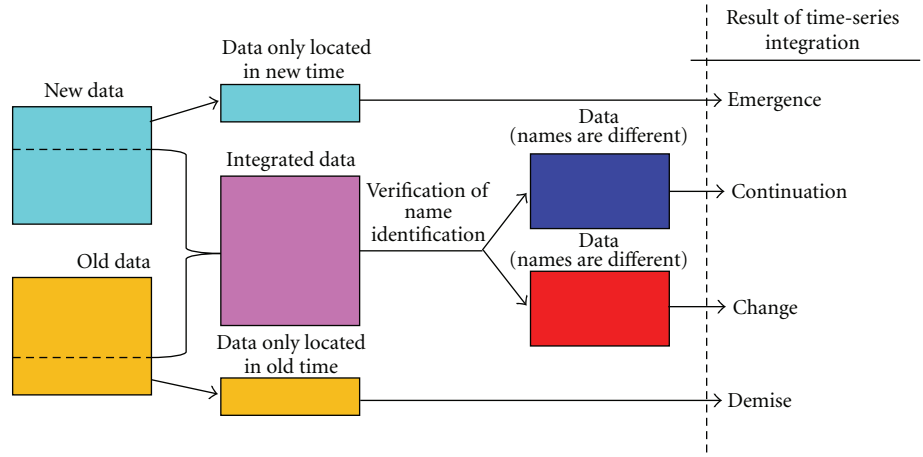

Figure 2: Processing flow of time-series integration by our method.

It was necessary to designate a minimum threshold for the similarity metric $S_{ij}^{(n)}$ experimentally. First, 3,000 shops and companies were randomly extracted from the 2005 Tokyo telephone directory, and then integrated with the shops and companies in their respective spaces from the 2000 directory. Next, the name similarities between shops and offices in the integrated dataset were calculated using the method above, using a comprehensive range of values for $S_{ij}^{(n)}$. We then compared these automated results with results obtained via manual processing that were verified as correct. Figure 4 shows the results of this comparison. Accordingly, a value of about 0.4 was determined to be optimal for the threshold of $S_{ij}^{(n)}$. Integrated data over the threshold $S_{ij}^{(n)}$ are considered to accord. As a result, we set the default value of $S_{ij}^{(n)}$ as 0.4 for our system.

3.4. Removal of Noise Words. Shop and company names may often contain frequently appearing words (FAWs), geographic names, and station names. Because of the confounding and pseudosimilar effects of these words and names, appropriate verification that similar names refer to the same tenant is difficult to achieve. Sagara and Kitsuregawa have also pointed out this difficulty in recognizing pure shop and company names using computers [26]. A method that can remove these so-called "noise words" from name information is necessary.

We solved this problem by creating dictionaries of noise words and using them to remove noise words from shop and company names prior to n-gram analysis. The FAW dictionary developed in this study was developed by applying an automated system of Japanese morphological analysis called "Chasen" [27] to tenant names that had been extracted from the 2005 residential maps and telephone directory covering the South Kanto region. Tenant names were divided into parts of speech by the Chasen, and these data were combined with manually culled FAWs to develop our library. Table 7 shows examples of FAWs taken from tenant names by automated morphological analysis of the Chasen. The library of geographic names was developed using the "Nihon Gyosei Kukaku Binran Data File" published by Nihon Kajo-Syuppan Corporation. The library of station names was developed

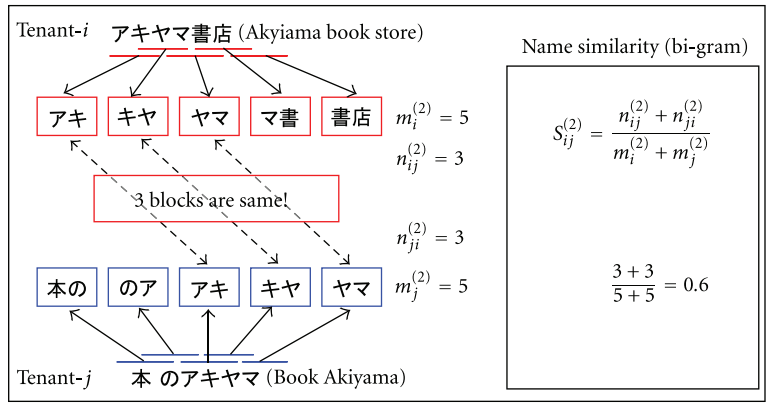

Figure 3: Calculation method for word similarity using a bigram.

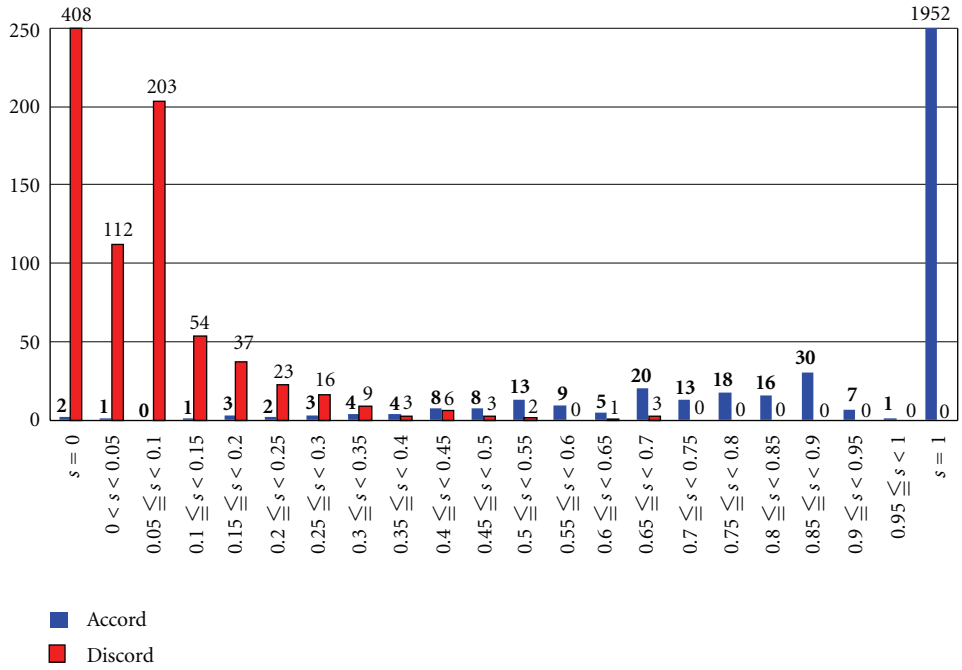

Figure 4: Distributions of bigram values in the case of accord and discord.

using railroad timetables of Japan. Table 8 shows some examples of noise words, and Table 9 shows the number of words present in each library.

Only those character strings that contain the geographic and station name structures depicted in Figure 5 are removed from shop and company names. The processing in Figure 5 is necessary because it decreases the risk that character strings which have no relation with geographic and station names might be removed from shop and company names. This risk increases to remove noise words in shop and company names to use only geographic and station names. We consider one geographic name, "Nakagawa (中川)", as an example. This name is common in Adachi ward in Tokyo Prefecture, where it refers to a geographic name. However, it is strongly expected that there will also be many shops and companies that contain "Nakagawa" even outside of the Nakagawa area, because it is a very popular family name in Japan: in fact, in the 2005 telephone directory, there are 311 shops and companies that do. Almost all of the shops and companies extracted were this nongeographical Nakagawa (Figure 6). Table 10 shows some examples of shop and company names containing an instance of Nakagawa unrelated to its geographic and station usages.

On the other hand, using the removal procedure in Figure 5 diminishes the risks associated with removing character strings unrelated to geographic and station names, that is, used for the n-gram similarity metric. Figure 7 shows the results of a search for shops and companies containing "Nakagawa" in their names using this rule. Two shops were found, and Table 11 shows that the "Nakagawa" in their names refers to the geographic Nakagawa.

Eventually, pure shop and company names remain after removing the various kinds of noise words through the above processing. This process is demonstrated in Figure 8.

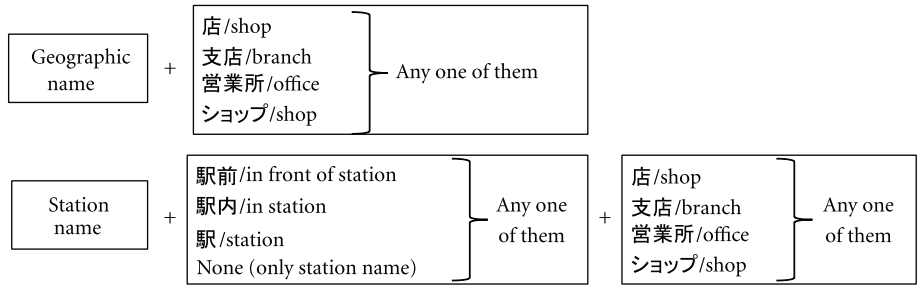

Figure 5: Character strings removed from shop and company names.

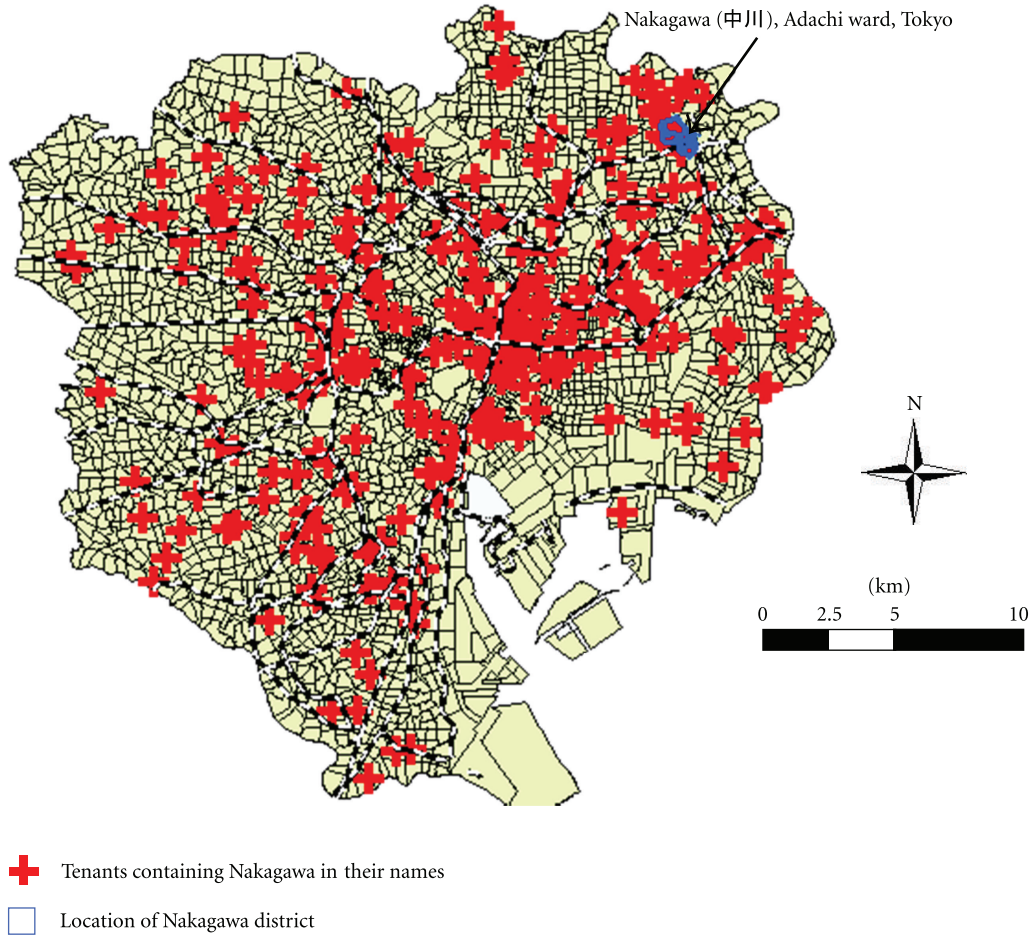

Figure 6: Locations of shops and companies in the 2005 telephone directory containing "Nakagawa (中川)" in Tokyo's 23 wards.

3.5. Removal of Local Frequently Appearing Words. Nonetheless, there are cases that cannot be processed well, even when all of the above methods and libraries are incorporated. This is because there are frequently appearing words that are eccentrically located both spatially and temporally. We refer to such words as "Local Frequently Appearing Words (LFAW)" in this study.

We explain about the LFAWs using three examples, depicted in Figure 9. "Shinjuku Nishiguchi" (the western exit of Shinjuku terminal) and "Yaesu-guchi" (Yaesu exit) are not geographic names. In addition, "Yaesu-guchi" is not a station name. However, there are many shops and companies whose names contain these character strings, because they are located in the western area of Shinjuku terminal or the eastern area of Tokyo terminal, respectively. These are examples of FAWs, which are eccentrically located in space: that is, they are concentrated only around a particular area. Figure 10 shows the locations of shops and companies whose names contain "Shinjuku Nishiguchi-ten" (Shinjuku terminal western exit shop) taken from the 2005 telephone directory. There are many data points in the western area outside Shinjuku terminal that fit this category.

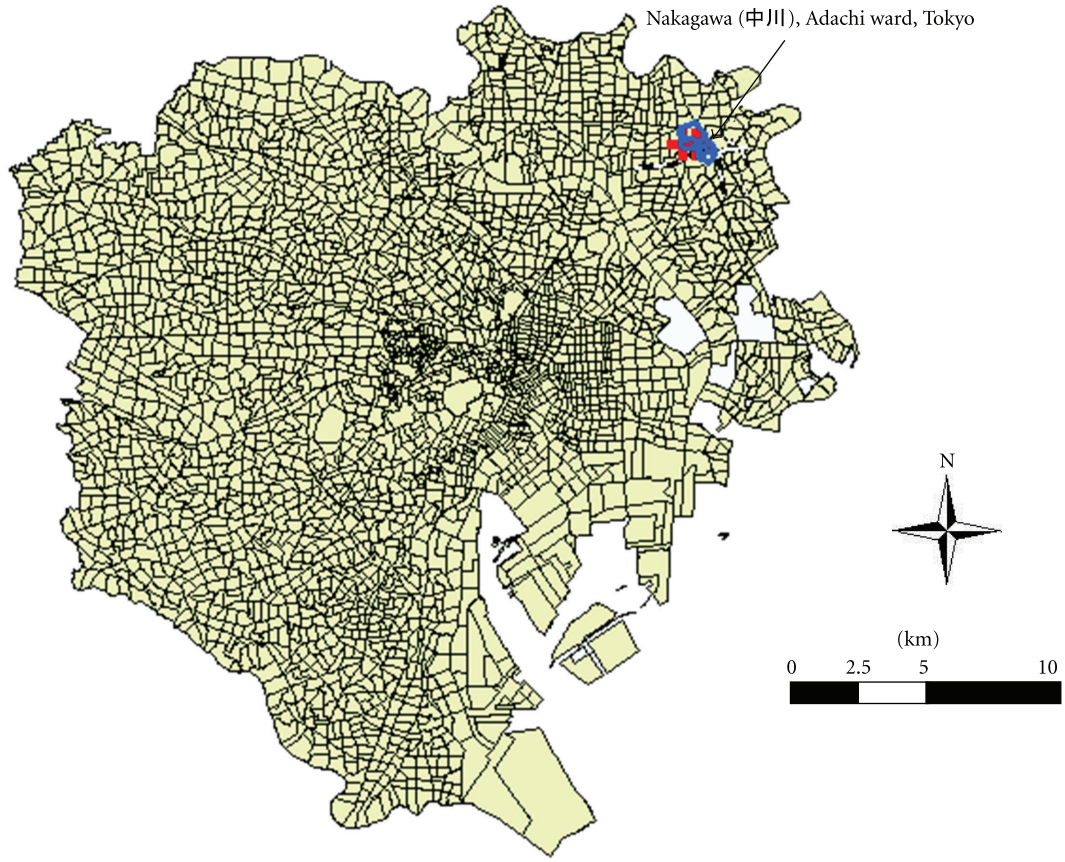

Figure 7: Search results of shops and companies containing "Nakagawa" as a geographic name.

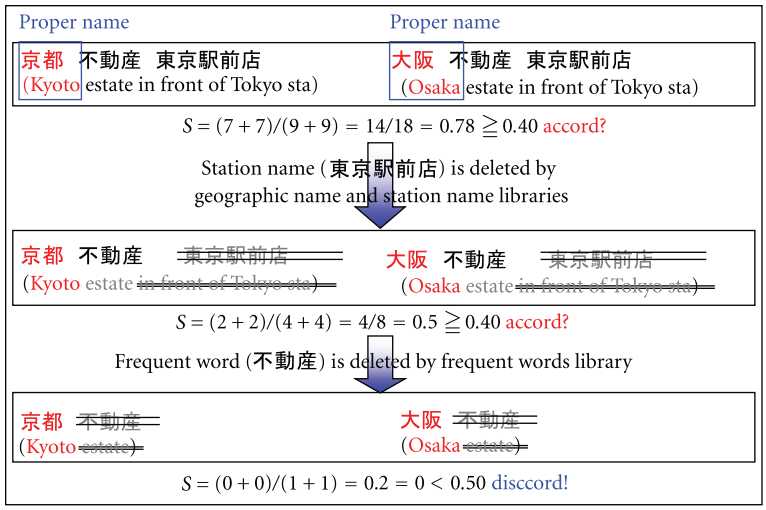

Figure 8: Accuracy improvement in bigram processing achieved by the removal of noise words.

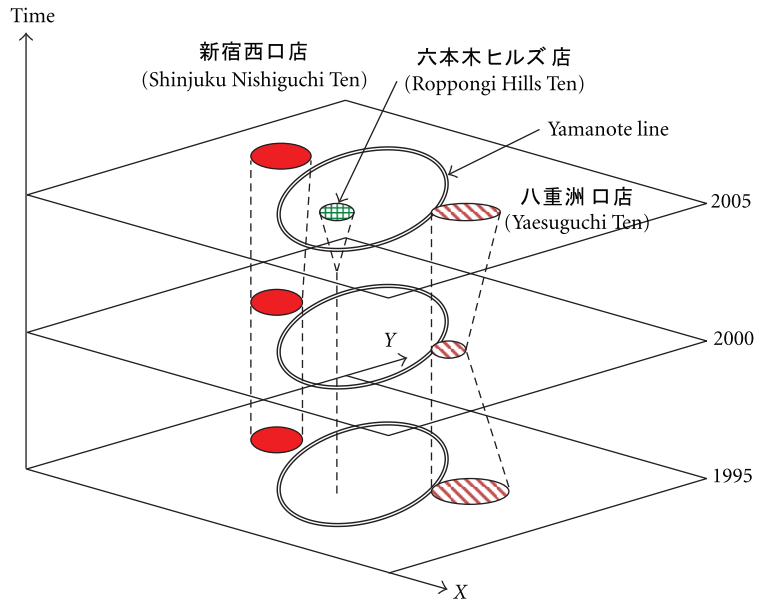

Figure 9: Image of LFAWs.

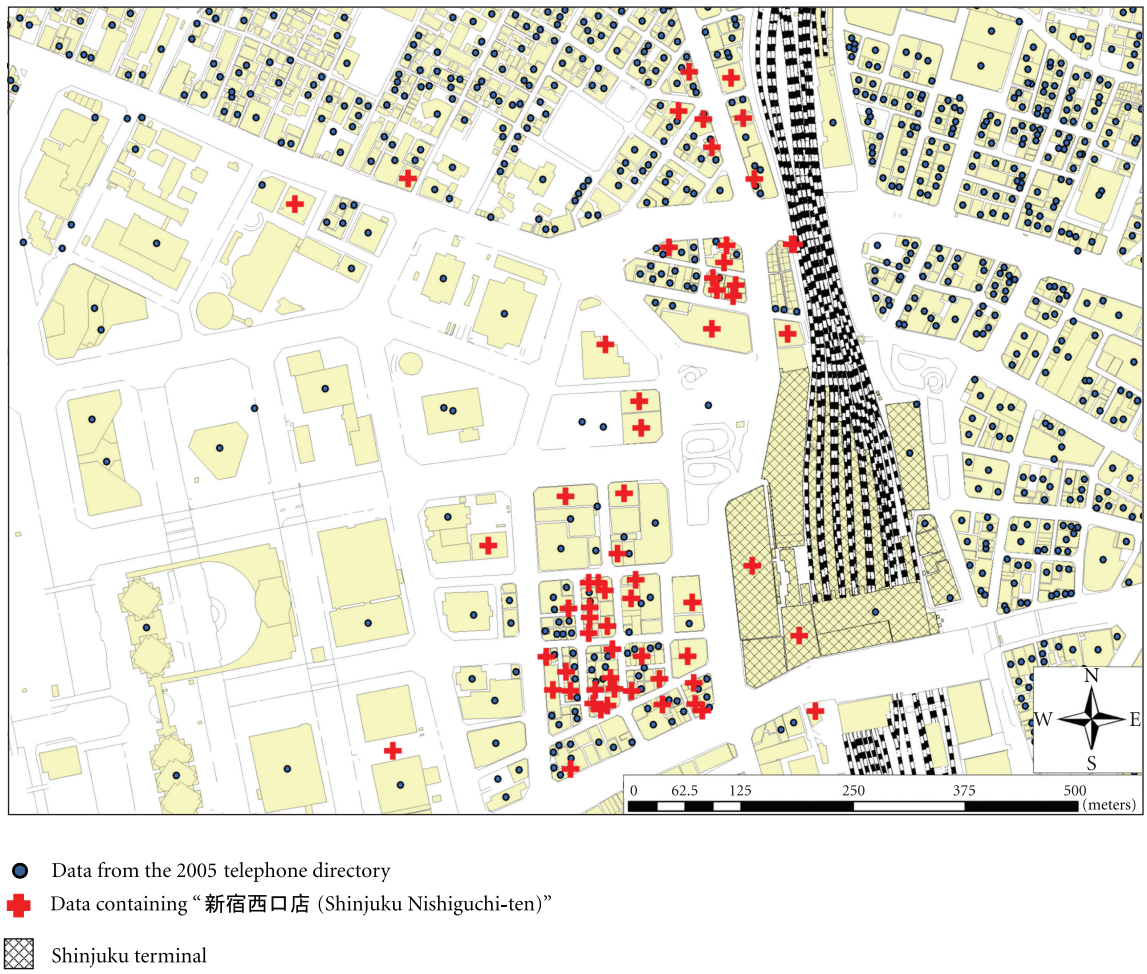

Figure 10: Data distribution of locations containing "Shinjuku Nishiguchi-ten" ("Nichiguchi-ten" means western exit shop/branch).

The list of LFAW (GridID: 13970035686)

Number of characters	LFAW list	Frequency	Examples	
			Names before removal of LFAW (red strings = LFAW)	Names after removal of LFAW
9	新宿西口国際通り店	2	福善新宿西口国際通り店	福善
8	新宿エステック店	3	ちよだ鮨新宿エステック店	ちよだ鮨
8	エステック情報店	3	そじ坊新宿エステック情報店	そじ坊新宿
7	ク新宿南口支店	2	ディック新宿南口支店	ディッ
6	ト新宿南口店	2	協和コンタクト新宿南口店	協和コンタク
6	新宿駅西口店	2	ミニストップ新宿駅西口店	ミニストップ
5	新宿南口店	3	山頭火新宿南口店	山頭火
4	新宿南口	12	天狗新宿南口	天狗
4	新宿西口	12	ボーダフォン新宿西口	ボーダフォン
4	新宿支社	3	ジャパンプロテック新宿支社	ジャパンプロテック
4	西口本店	2	ヨドバシカメラ西口本店	ヨドバシカメラ
4	東京支社	4	近畿設備東京支社	近畿設備

- The search area is 3/1000° square (the SW edge is N35.685, E139.699 and the NE edge is N35.688, E139.702.).
- The names before LFAW removal have already had other noise words removed.

Figure 11: Examples of LFAWs and their removal from shop and company names in one grid covering an area west of Shinjuku terminal (taken from the 2005 telephone directory).

On the other hand, "Roppongi Hills" (one of the largest and most famous skyscraper complexes in Tokyo and Japan) in Figure 9 is an example of an FAW, which is eccentrically-located not only spatially but also temporally. There are almost no shop or company names from before 2003 containing "Roppongi Hills," because Roppongi Hills was only opened in that year. There are 105 shops and companies in 2005 and 141 in 2009 that contain "Roppongi Hills" in their names: however, there was zero such shops in the 2000 Tokyo telephone directory, as Roppongi Hills opened only in 2003. Thus, a method to remove these kinds of LFAWs was necessary.

We constructed grids measuring millidegree square along longitude and latitude, and all source data were allocated to this grid. Frequently appearing character strings were searched for in the shop and company names within each grid using the n-gram method. For each grid, n-grams created strings measuring from $n = 4$ to $n = 9$ based on both the shop and company names within the targeted grid itself and the neighboring 8 grids on all sides. It was necessary to search in the neighboring grids as well, so that shops or companies located near the grid borders could be incorporated into the LFAW identification process. Finally, the identified LFAWs were removed from the shop and company names in each grid. For our purposes, LFAWs are only 4- to 9-gram-constructed strings that appear multiple times, and whose endings comprise "店/ten" (shop), "支社/sisya" (branch), "営業所/eigyosyo" (office), "東口/higashiguchi" (eastern exit), "西口/nishiguchi" (western exit), "南口/Minamiguchi" (southern exit), and "北口/kitaguchi" (northern exit). Also, long LFAWs are removed earlier than short LFAWs. In other words, those LFAWs created by the 9-gram are removed first, those created by the 8-gram next, and so on through the 4-gram.

To this effect, Figure 11 shows the results from these example LFAWs and their removal in one grid west of Shinjuku terminal. Almost all of the shop and company names were processed adequately. In addition, even when

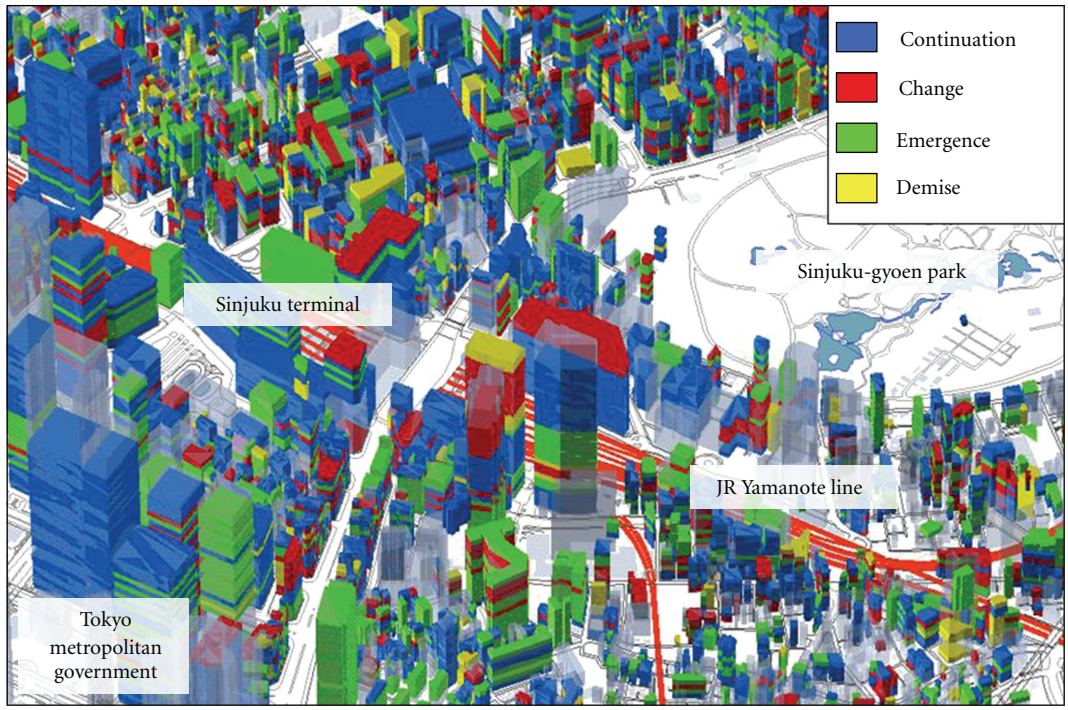

Figure 12: Time-series 3D map of tenants around Shinjuku terminal.

Table 9: The number of words in each noise word dictionary.

Regions of Japan	Number of words		
	Geographic names	Station names	FAW
Hokkaido	48570	1519	963
Tohoku	135436	1314	
North Kanto	12224	678	
South Kanto	35869	2625	
Koshinetsu	16677	785	
Hokuriku	14248	721	
Tokai	76361	2008	
Kinki	66681	2290	
Chugoku	21554	1081	
Shikoku	17303	580	
Kyusyu	28868	1717	
Okinawa	1835	19	

Each region contains the following prefectures.
- Hokkaido: Hokkaido
- Tohoku: Aomori, Iwate, Akita, Miyagi, Yamagata, and Fukushima
- North Kanto: Ibaraki, Tochigi, and Gunma
- South Kanto: Saitama, Chiba, Tokyo, and Kanagawa
- Koshinetsu: Niigata, Nagano, and Yamanashi
- Hokuriku: Toyama, Ishikawa, and Fukui
- Tokai: Shizuoka, Aichi, Gifu, and Mie
- Kinki: Siga, Kyoto, Nara, Wakayama, Osaka, and Hyogo
- Chugoku: Tottori, Shimane, Okayama, Hiroshima, and Yamaguchi
- Shikoku: Kagawa, Tokushima, Kochi, and Ehime
- Kyusyu: Fukuoka, Nagasaki, Saga, Oita, Kumamoto, Miyazaki, and Kagoshima
- Okinawa: Okinawa

parts of noise words remained in the pure shop or company names, or when parts of the pure name were erroneously removed, the effects can be largely ignored because the n-gram can still calculate name similarity effectively as in Figure 3.

Table 10: Examples of shop and company names containing Nakagawa unrelated by geography or station name.

Names	
In Japanese	In English
株式会社中川印刷所	Nakagawa Printing Press Co., Ltd.
中川金属株式会社	Nakagawa Metal Co., Ltd.
中川歯科クリニック	Nakagawa Dental Clinic
中川屋カレーうどん	Nakagawa-ya Curry Udon

Table 11: Shop names of search results from Figure 7.

Names	
In Japanese	In English
あいちや中川店	Aichiya Nakagawa shop
西沢薬局中川支店	Nishizawa pharmacy, Nakagawa branch

Tables 12 and 13 show some example test results of our method for removing noise words. For telephone directory data (Table 12), there are 654 shops and companies from which noise words should be removed, out of 1000 extracted by random sampling. 92.4% of these 654 shops and companies had their noise words successfully removed without damaging the character strings of pure names. For web information (Table 13), the fluctuation of shop and company names seems larger than in the telephone

Table 12: Processing accuracy of removal of noise words (Data consists of 1000 samples extracted randomly from the 2005 Tokyo prefecture telephone directory).

Number of samples	1000						
Is it necessary to remove noise words from names, as determined by a manual check?	Yes: 654				No: 346		
Can we get the same result as manual processing using the FAW dictionary?	**Yes: 513**		No: 141				
Can we get the same result as manual processing using the dictionary of geographic names and station names?		**Yes: 70**		No: 71			
Can we get the same result as manual processing after LFAW removal?			**Yes:11**	**No:60**			
Do pure names remain after all noise word removal processing?					**Yes: 330**	**No: 16**	**Sum total**
Number of data processed successfully	513	70	11	0	330	0	924
Processing accuracy (%)							92.40

Table 13: Processing accuracy of removal of noise words (Data consists of 1000 samples extracted randomly from web data using the Hot Pepper API from within Tokyo prefecture).

Number of samples	1000						
Is it necessary to remove noise words from names, as determined by a manual check?	Yes: 545				No: 455		
Can we get the same result as manual processing using the FAW dictionary?	**Yes: 67**		No: 478				
Can we get the same result as manual processing using the dictionary of geographic names and station names?		**Yes: 237**		No: 241			
Can we get the same result as manual processing after LFAW removal?			**Yes:81**	**No:160**			
Do pure names remain after all noise word removal processing?					**Yes: 409**	**No: 46**	**Sum total**
Number of data processed successfully	67	237	81	0	409	0	794
Processing accuracy (%)							79.40

"Hot Pepper" is a famous free coupon magazine in Japan, produced by Recruit Co., Ltd. Using the Hot Pepper API, we can collect information about many kinds of shops, companies, restaurants, and so forth.

directory, with 79.4% of the locations having their noise words removed successfully. In addition, the FAW dictionary exerts the largest discriminative effect within each test.

There are cases where noise words still remain partly in shop and company names or where important character strings are erroneously removed after the LFAW processing. These are indicated by the blue numbers in Table 12 (76 shops and companies: 7.6%) and Table 13 (206 shops and companies: 20.6%) and the blue character strings in the "Name after removal of LFAW" row within Figure 11.

Table 14: Sample areas and their numbers of data.

Sample areas		The number of data in 2005		in 2000	
Locations of area	Area types	Z	T	Z	T
Part of Kabukicho, Shinjuku-ku, Tokyo	Bustling shopping area in city center	191	88	210	94
Nodai dori shopping street, Setagaya-ku, Tokyo	Old shopping street around train station	276	130	247	141
Vicinity of Amatsu port, Kamogawa city, Chiba	Port town	121	82	136	94
Nakanogo and Kashidate districts, Hachijo island, Tokyo	Settlements in an isolated island	140	105	144	121
	Sum total	728	405	737	450

"Z" in number of data denotes residential map tenant data.
"T" in number of data denotes telephone directory data.

Table 15: Processing accuracy of time-series integration for residential map tenant data.

		System results						
		Sum total	Co	Ch	Em	De	FSI	Accuracy (%)
Manual results	Continuation	**467**	**443**	4	22	0	2	**94.86**
	Change	**129**	4	**121**	8	0	0	**93.80**
	Emergence	**132**	4	2	**126**	0	0	**95.45**
	Demise	**141**	8	3	0	**130**	0	**92.20**
						Sum total		**94.36**

"Co": continuation, "Ch": change, "Em": emergence, and "De": demise
"FSI" means failure of spatial integration.

However, these effects are negligibly small, because n-gram processing can nonetheless verify name similarity despite the incompleteness of the pure character string. Tables 15, 16, and 17 in Section 4 will demonstrate that our method can process data with sufficiently high accuracy.

4. Processing Accuracy

So far, we have developed a method for removing various kinds of noise words from shop and company names, and one for verifying that differing names may refer to the same tenant, by calculating the name similarity. In this section, the processing accuracy of our system achieved by these methods is discussed.

We compared the results of time-series data produced by our system with results created manually (and verified for correctness) in some sample areas in the South Kanto region of Japan. Input data for this verification of identity were taken from the telephone directories and digital residential maps as described in Section 1.3. Table 14 shows our sample areas: two each of urban and rural areas.

First, telephone directory and residential map tenant data from 2000 were integrated spatiotemporally with the same data from 2005 over the whole South Kanto region. Then, sample data were extracted from the results of time-series integration. Finally, these automated results were compared with results manually obtained and verified as correct.

Tables 15 and 16 show the processing accuracy achieved by our system's time-series integration. Each table begins with the manually verified total of all the time-series changes observed in each of the sample areas. The system accomplished a processing accuracy of 94.36% (820/869) in integrating the old and new residential map tenant data spatiotemporally (Table 15), and one of 95.22% (478/502) in integrating the old and new telephone directory data (Table 16). The reason why the sum totals here are discordant with their respective sum totals in Table 14 is because the "Demise" results are counted instead of obscured by subtraction. That is, the sum total in Table 15 not including "Demise" data is the same as the sum total of the number of tenants in the 2005 residential maps. The most remarkable and salient point from Tables 15 and 16 is the high accuracy achieved for continuation and change results. We could not have acquired such high values without not only accurate spatial integration but also a robust method for identifying differing names as referring to the same tenant. We demonstrate that the Japanese language processing methodology introduced in this paper is effective for the realization of time-series integration.

In addition, Table 17 shows the processing accuracy for the residential map tenant data in each sample area. For example, in the Continuation column of "Kabukicho," "79 (87)" means that 87 data points were judged as "Continuation" through manual analysis, and 79 out of those 87 were judged to be the same time-integration category by our system. Processing accuracies in urban areas were slightly lower than in rural areas, because of high density of shops and companies and frequent transfers of them. Processing accuracies in rural areas were almost 100%.

It has been shown in Tables 15, 16, and 17 that there is about a 5% error rate when creating time-series data manually. Compared to this rate, the processing accuracy of

Table 16: Processing accuracy of time-series integration for telephone directory data.

		System results						
		Sum total	Co	Ch	Em	De	Sim	Accuracy (%)
Manual results	Continuation	**307**	**293**	4	10	0	0	**95.44**
	Change	**46**	1	**43**	2	0	0	**93.48**
	Emergence	**52**	0	3	**49**	0	0	**94.23**
	Demise	**97**	2	2	0	**93**	0	**95.88**
							Sum total	**95.22**

Table 17: Comparison of processing accuracy in each sample area.

Sample area	Number of data		Processing accuracy of time-series integration					Accuracy (%)
	in 2005	in 2000	Co	Ch	Em	De	FSI	
Part of Kabukicho, Shinjuku-ku, Tokyo	191	210	79(87)	67(74)	28(30)	44(49)	2	**94.58**
Nodai-dori shopping street, Setagaya-ku, Tokyo	276	247	137(154)	48(51)	67(71)	38(42)	0	**92.45**
Vicinity of Amatsu port, Kamogawa city, Chiba	121	136	116(116)	0(1)	4(4)	19(19)	0	**99.29**
Nakanogo and Kashidate districts, Hachijo island, Tokyo	140	144	106(110)	2(3)	27(27)	29(31)	0	**97.66**

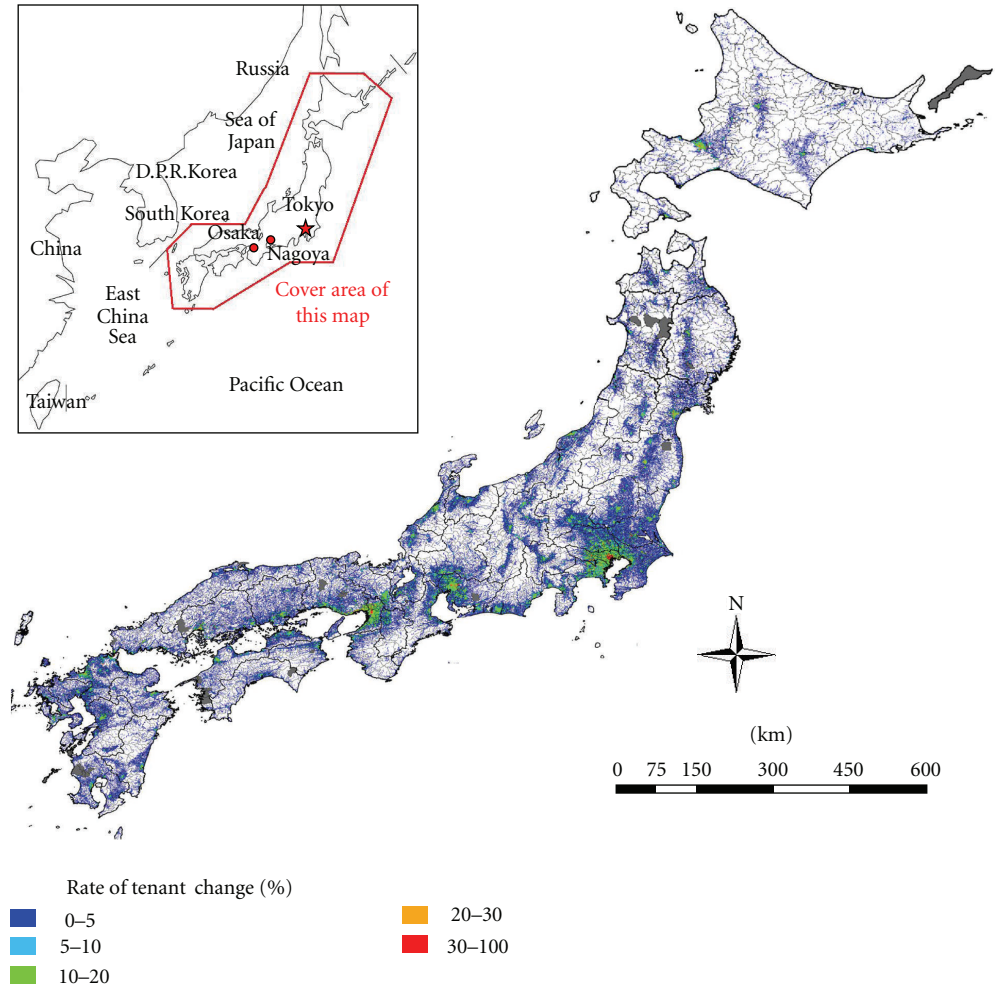

Figure 13: Grid map of rate of tenant change all over Japan (1 km square grid).

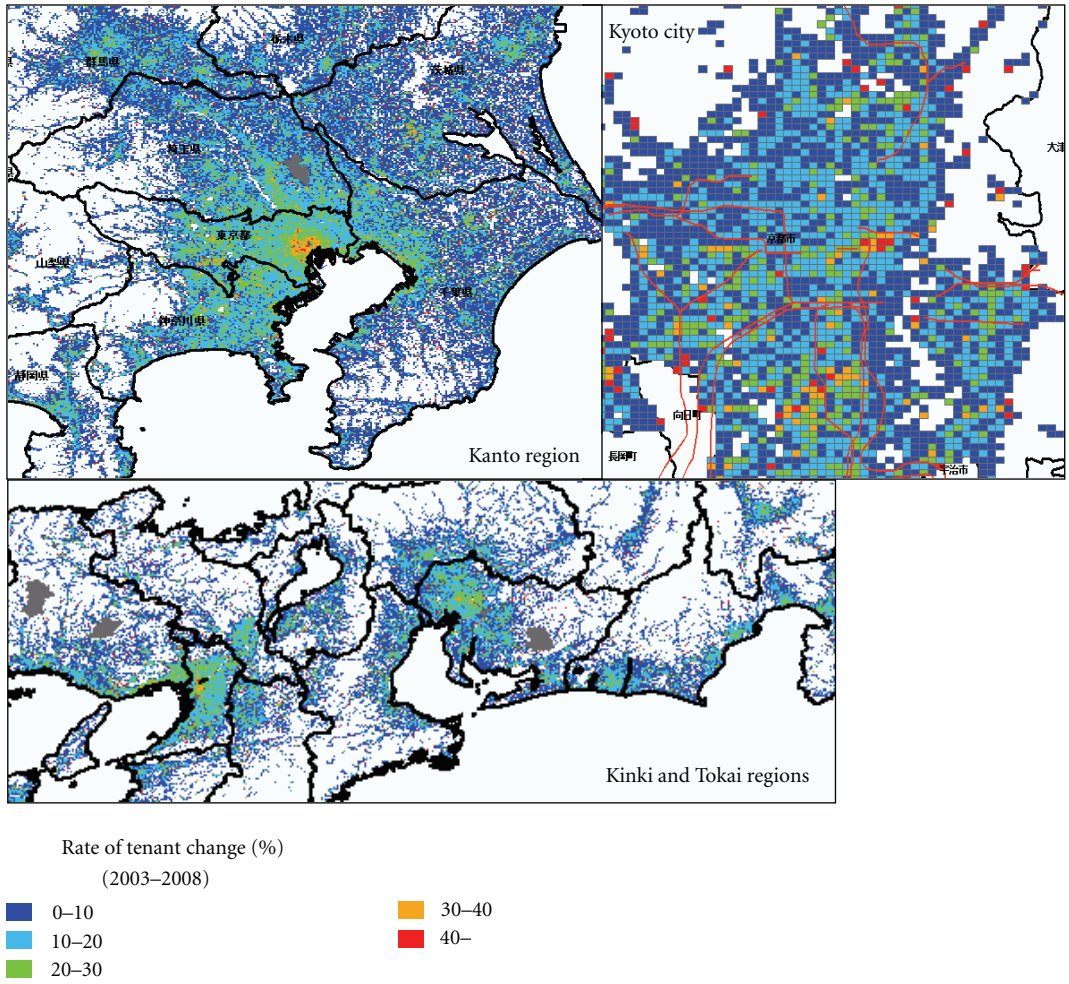

Figure 14: Grid map of rate of tenant change in parts of Japan (500 m square grid).

our system is certainly practical and robust, considering the inevitable human error and large amounts of labor and time necessary when performing such work manually.

The high processing accuracy observed with our system was achieved not only with accurate spatiotemporal integration—the method we developed, which identifies two different names as referring to the same tenant by calculating of word similarity, was essential. The results detailed in this section demonstrate that our method for name identification discussed in this paper performs at a reliable level.

5. Examples of Data Graphics Developed Using Our System

In this section, we will briefly discuss some examples and applications of detailed time-series datasets that can be developed by our system.

Figure 12 shows a 3D time-series map of tenant changes around Shinjuku terminal between 2000 and 2005. This map was constructed so as to integrate the residential map time-series datasets with the respective telephone ones for the years 2000 and 2005. It is possible to find buildings that were newly built between 2000 and 2005 by searching for buildings where all tenants are categorized as "Emergence," and conversely, to find vacant sites or sites under construction by searching for buildings where all tenants are categorized as "Demise." In addition, we can easily see that many of the "Change" tenants are (in 2005) located in low floors around Shinjuku terminal.

Figure 13 shows a grid map (500 m square length) of various "Change" rates based on the results of time-series integration of residential map tenant data from 2003 and 2008 from all over Japan. This is calculated as the number of "Change" tenants divided by the total number of tenants. It is readily apparent from Figure 14 that grids with high "Change" rates are located in urban areas: this may be expected, since competition among shops and companies is usually intense in such areas. In addition, it is interesting to be able to monitor the variability of the "Change" rate across many different areas in the same city.

This is the first instance where such a detailed time-series dataset with such homogenous resolution over this broad of an area has been realized in Japan. This kind of data can make a valuable contribution in solving the problems encountered in previous studies, as introduced in Section 1.

6. Conclusion

In this paper, we discussed a method for identifying Japanese names, by quantitatively analyzing their true, "pure" similarities while ignoring pseudosimilar "noise words" within them. The most remarkable achievement of this study was its removal of eccentricallylocated LFAW located both spatially and temporally by an *n*-gram-adapted methodology. This novel approach integrates knowledge bases from both linguistics and spatial information science. In addition, we can further conjecture that this study is predictive of how the demands for natural language processing will increase more and more in the fields of spatial information science and geography.

There are some future challenges to improve the identification of Japanese words. One challenge is to develop an environment that can convert effortlessly between kanji, hiragana, and katakana. Mutual conversion between hiragana and katakana is very easy because both sets of characters comprise the same set of phonograms. However, it seems difficult to convert kanji directly into hiragana or katakana because kanji are ideograms. In addition, almost all kanji in Japanese have multiple kinds of pronunciation. The development of a method to accurately and robustly convert kanji into hiragana or katakana is one of the most important tasks facing our research. Another important challenge is that of converting loanwords into katakana. We have already realized a simplified system that can do this. However, the processing accuracy of this system is inadequate, with this system converting only some English and French words into katakana precisely. Both are very difficult challenges, yet nonetheless very interesting and exciting directions for future research.

Acknowledgments

The authors were given the digital telephone directory by ZENRIN CO., LTD (Telepoint Pack!) and NTT Business Information Service, Inc. (Town Page Databese) and the digital residential maps by ZENRIN CO., LTD (Zmap TOWN II). Publication of this paper was supported by Earth Observation Data Integration and Fusion Research Institute (EDITORIA). They would like to thank ZENRIN CO., LTD, NTT Business Information Service, Inc., and EDITORIA for their contribution.

References

[1] Y. Akiyama and R. Shibasaki, "Development of detailed spatio-temporal urban data through the integration of digital maps and yellow page data and feasibility study as complementary data for existing statistical information," in *Proceedings of the Computers in Urban Planning and Urban Management (CUPUM '09)*, 187, 2009.

[2] Y. Akiyama, T. Shibuki, and R. Shibasaki, "Development of three dimensional monitoring dataset for tenants variations in broad urban area by spatio-temporal integrating digital house maps and yellow page data," in *Proceedings of the 4th International Conference on Intelligent Environments (IE '08)*, 2008.

[3] T. Ato, K. Omura, T. Arata, and S. Hujii, "The stagnation of commercial accumulation districts in front of the stations in the suburbs of the Tokyo metropolitan area: a study of honatsugi and odawara," *City Planning Review*, vol. 41, no. 3, pp. 1037–1042, 2006.

[4] H. Ai, Y. Sadahiro, and Y. Asami, "Spatio-temporal analysis of building location and building use in middle scale commercial accumulation districts," *City Planning Review*, vol. 43, no. 3, pp. 103–108, 2008.

[5] K. Ito and H. Magaribuchi, "Method for making spatio-temporal data from accumulated information: using the identification by resolving geometric and non-geometric ambiguity," in *Proceedings of the Geographic Information Systems Association*, vol. 10, pp. 147–150, 2001.

[6] R. Florian, H. Hassan, A. Ittycheriah et al., "A statistical model for multilingual entity detection and tracking," in *Proceedings of the Human Language Technologies Conference (HLT-NAACL '04)*, pp. 1–8, May 2004.

[7] Q. Tri Tran, T. X. Thao Pham, Q. Hung Ngo, D. Dinh, and N. Collier, "Named entity recognition in Vietnamese documents," *Progress in Informatics*, no. 4, pp. 5–13, 2007.

[8] E. F. Tjong Kim Sang and F. D. Meulder, "Introduction to the CoNLL-2003 Shared task: language-independent named entity recognition," in *Proceedings of the 7th Conference on Natural Language Learning (HLT-NAACL '03)*, vol. 4, pp. 142–147, 2003.

[9] R. Florian, A. Ittycheriah, H. Jing, and T. Zhang, "Named entity recognition through classifier combination," in *Proceedings of the 7th Conference on Natural Language Learning at (HLT-NAACL '03)*, vol. 4, pp. 168–171, 2003.

[10] H. L. Chieu and H. T. Ng, "Named entity recognition: a maximum entropy approach using global information," in *Proceedings of the 19th International Conference on Computational Linguistics*, vol. 1, pp. 1–7, 2002.

[11] R. Steinberger and B. Pouliquen, "Cross-lingual named entity recognition," *Lingvisticae Investigationes*, vol. 30, no. 1, pp. 135–162, 2007.

[12] T. Bogers, *Dutch named entity recognition: optimizing features, algorithms, and output*, Ph.D. thesis, University of Van Tilburg, 2004.

[13] C. Sporleder, M. V. Erp, T. Porcelijn, A. V. Bosch, and P. Arntzen, "Identifying named entities in text databases from the natural history domain," in *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pp. 1742–1745, 2006.

[14] S. Sato, M. Harada, and K. Kazama, "Measuring similarity among information sources by comparing string frequency distributions," *Information Processing Society of Japan Digital Document*, vol. 2002, no. 28, pp. 119–126, 2002.

[15] T. Kawakami and H. Suzuki, "A calculation of word similarity using decision list," *IPSJ SIG Technical Report*, vol. 2006, no. 94, pp. 85–90, 2006.

[16] K. Mishina, S. Tsuchita, S. Kurokawa, and R. Huji, "An emotion similarity calculation using N-gram frequency," *IEICE Technical Report*, vol. 107, no. 158, pp. 37–42, 2007, NLC2007-7.

[17] D. Cali, A. Condorelli, S. Papa, M. Rata, and L. Zagarella, "Improving intelligence through use of natural language processing. A comparison between NLP interfaces and traditional visual GIS interfaces," *Procedia Computer Science*, vol. 5, pp. 920–925, 2011.

[18] B. Bitters, "Geospatial reasoning in a natural language processing (NLP) environment," in *Proceedings of the 25th International Cartographic Conference, CO-253*, July 2011.

[19] S. Miyagawa, "The Japanese Language," MIT JP NET, 2011, http://web.mit.edu/jpnet/articles/JapaneseLanguage.html.
[20] D. Klein, J. Smarr, H. Nguyen, and C. D. Manning, "Named entity recognition with character-level models," in *Proceedings of the 7th Conference on Natural Language Learning (HLT-NAACL '03)*, vol. 4, pp. 180–183, 2003.
[21] S. Kuno, *The Structure of the Japanese Language. Current Studies in Linguistics*, MIT Press, 1 edition, 1973.
[22] C. E. Shannon, *A Mathematical Theory of Communication*, University of Illinois Press, 1948.
[23] M. Kondo, *An Analysis of Japanese Classical Literature Using Character-Based N-Gram Model*, vol. 29, Chiba University, Zinbun Kenkyu, 2000.
[24] T. Odaka, T. Murata, J. Gao et al., "A proposal on student report scoring system using N-gram text analysis method," *Journal of Institute of Electronics, Information, and Communication Engineers*, vol. 86, no. 9, pp. 702–705, 2003.
[25] J. B. Marino, R. E. Banchs, J. M. Crego et al., "N-gram-based machine translation," *Computational Linguistics*, vol. 32, no. 4, pp. 527–549, 2006.
[26] T. Sagara and M. Kitsuregawa, "Cleaning shop names by its location information for shop information retrieval from the web," *Journal of Institute of Electronics, Information, and Communication Engineers*, vol. 91, no. 3, pp. 531–537, 2008.
[27] Chasen legacy—an old morphological analyzer, http://chasen-legacy.sourceforge.jp/.