Hot Events Analysis based on Japanese Corpora processing

WANG Tong[1,a], YI Mianzhu[1,b]

[1]Luoyang Campus, PLA Information Engineering University, Guangwen Street, Luoyang, China

[a] 463906155@qq.com, [b] 1197751829@qq.com

Keywords: Japanese name recognition, Map database, Hot events analysis, Asahi Shimbun, Yasukuni Shrine

Abstract. Lacking of study on sufficient foreign literature and data, domestic researches on hot events between China and Japan are one-sided due to the gap between languages. In order to analysis hot events more objectively by Japanese corpus processing and knowledge mining, this paper puts forward a method of name recognition and reference resolution based on surname dictionary and syntax rules. What is more, a characters detection model of hot events is designed and the knowledge graph of core characters is showed on map database. Finally, we tested the Japanese name recognition method on Asahi Shimbun corpus, of which the accuracy was 90.07% and the recall rate was 91.15%.

1. Introduction

The diplomatic relationship between China and Japan has always been one of the focuses on China's diplomatic chessboard, due to the neighbouring position and long historical origin. Therefore, the Sino Japanese hot events are of great significance on catching an objective view of current relationship and making scientific predictions. Barriers, such as the language gap and the inadequate data, lead to the biased results, which are not scientific and convincing at all. In the era of big data, there is no doubt that Japanese hot event news corpus is precious materials and basic resources for research and analysis.

Name recognition is an effective tool in text analysis, with which can we find out the high frequency characters, one of the essential elements in the text and establish solid foundation for events analysis. Till now, domestic name recognition are focus on English names and Chinese names, such methods are not suitable for Japanese issues. ZHANG Han [1] proposed a method of Japanese name recognition in syntagma-segments. 平田亜衣 [2] extracted the names from news corpus with boundary tags in his work 《様々なジャンルのテキストに対する固有表現認識の分析》. 浅原正幸 and 松本裕治 [3] designed a morphological parser to locate and determine the name boundary in their work《日本語固有表現抽出におけるわかち書き問題の解決》. Reviewing these researches, name abbreviation and brief reference such as surname followed by title were ignored, leading the results incomprehensive. It is undisputed that name frequency is a pivotal parameter in event analysis. In order to solve the above problem and increase the recall rate of Japanese name recognition, our work can be summarized as following. Firstly, crawl Asahi Shimbun news and build a Japanese news corpus. Secondly, according to the attached tags and syntactic rules, we present a Japanese name recognition method from raw corpus and design a detector of high frequency characters in hot event with anaphora regression. Thirdly, construct a Hash net of high frequency characters and a knowledge map using neo4j graph database as well. Finally, take the Yasukuni Shrine event as a sample to test the validity of our research.

2. Japanese name recognition from news corpus

2.1. Construction of Japanese news corpus

Corpus is a collection of corpora with a certain structure, representative, and can be retrieved by computer programs for one or more applications. According to different criterion, it can be divided into monolingual corpus and multilingual corpus, homogeneous corpus and heterogeneous corpus, systematic corpus and specific corpus, spoken language corpora and text corpus.

This paper selects Asahi Shimbun, one of the three most comprehensive Japanese newspapers, as the source of language acquisition. Asahi Shimbun has been published for more than 120 years, which has great influence in Japan and has a large readership. Taking account of event pertinence, data consistency and content comprehensiveness, we build a monolingual, homogeneous full text Japanese news corpus in order to feed the research demands. The steps of corpus construction are as follows (take the Yasukuni Shrine event as an example):

1.Set Yasukuni Shrine as the keyword and set the limited time interval from 1st January 2000 to 31st December 2015, use web collection tools and get multi-threaded parallel access to Asahi Shimbun website target news page.

2. Write regular expressions according to web page format and take the label impurities away. As a result, each text is separate storage with marked title, time and context.

3. Remove duplications according to title and publication time because the same news agency will not publish the same headlines on the same day. Finally obtain a Yasukuni Shrine corpus of 7716 news related.

2.2. Detector of high frequency characters

2.2.1. Structure of the detector

Through previous research alongside the analysis of the news corpus we built above, difficulty of Japanese name recognition falls into the following categories：

1. Compared with the Latin language represented by the English language, there is no obvious word boundary marker in Japanese text, which leads to the segmentation ambiguity.

2. There are about over 100 thousand surnames in Japanese name, so that the conventional dictionary storage structure occupies a large amount of memory space and the matching time complexity is intolerable.

3. The length of Japanese names varies from 2 characters to 8 characters, which increases the difficulty of defining the name boundary.

Aiming at solving the above difficulties, we design the detector with three principles, to optimize the storage structure, to specify boundary rule and to improve the matching efficiency. The structure of the detector can be shown as figure 1.

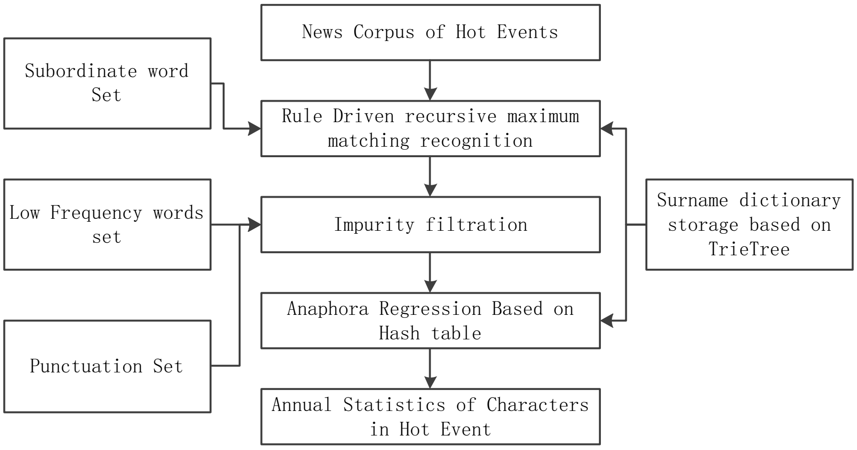
News Corpus of Hot Events
Subordinate word Set
Rule Driven recursive maximum matching recognition
Low Frequency words set
Impurity filtration
Surname dictionary storage based on TrieTree
Anaphora Regression Based on Hash table
Punctuation Set
Annual Statistics of Characters in Hot Event

Figure 1 System flow chart of high frequency characters detector.

2.2.2. Surname dictionary storage based on TrieTree

The surnames dictionary we used contains 7938 common terms. TrieTree is chosen as the storage structure due to the its large size and prefix redundancy. TrieTree is a kind of multi tree, also known as dictionary tree, is widely used in dictionary storage. The core idea of TrieTree is using the shared prefix to make associative storage of the strings. In other words, the same prefix is stored only once, resulting in the maximum compression of the storage space and the improvement of query efficiency as well. For example, the storage of a surname dictionary {井野，井野邊，井野川，井口，柳町，柳原，穀田，穀田部，穀田貝，穀田川} is shown as figure 2.

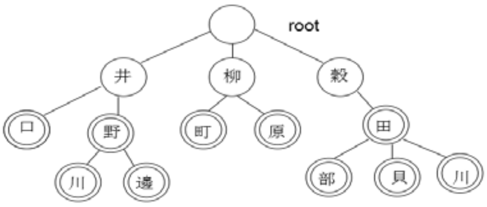

Figure 2 Storage structure of surname dictionary based on TrieTree.

2.2.3. Rule Driven recursive maximum matching recognition algorithm

Research on news corpus shows that the expression and contextual features of Japanese names come in several varieties: Firstly, surname is in front of name. Secondly, Japanese belongs to the adherent language family, as an independent word, the name must be accompanied by the attachment of the subordinate words. For example, ……三島由紀夫が……“三島由紀夫”is a name, “が”is a subordinate word. Thirdly, with a title or an appellation directly linked to a name. For example, ……鈴木善幸首相……“鈴木善幸” is a name，“首相” is a title means prime minister. Another example, ……中島順子先生……“中島順子” is a name，“先生” is a appellation means sir. Fourthly, after the complete expression of a name, it is often used as the abbreviation of surname followed with title in the news corpus.

To sum up, Japanese name chunk can be formalized as followed:

<Surname>+[<name>]+[<title>]|[< appellation >]+<subordinate word>

The rule driven recursive forward maximum matching algorithm is illustrated as 4 steps:

1.When the text matches a node mark as the complete surname, continue traversing a finite depth downwards, if the node mark is a complete surname too, outputs it directly; if the mark is a non complete surname, back to the nearest node marked as a complete surname, search from this node, until a node flag is a complete surname, and its lower parameter deep sub node flag are not. And it is the dynamic longest matching of surnames.

2.With the certain surname, continue scanning the text until find an element in the subordinate word set, cut the segment as a candidate Japanese name chunk.

3.Mark the length of the surname for subsequent impurity filtration and anaphora regression.

2.2.4. Impurity filtration and anaphora regression

Through the analysis of candidate Japanese name chunk identified by the test set, it is found that the impurities mainly include:

1.Places wrongly recognized as names, for example “秋田市の”、“吉野町の”.

2.Overlapping ambiguity caused wrongly recognized, for example, since “三角”is a term in the surname dictionary, “三角形の”is wrongly recognized as a name chunk.

3.Combinatorial ambiguity caused wrongly recognized, for example, since “公文”is a term in the surname dictionary, “公文署名式や” is wrongly recognized as a name chunk.

4.A surname shows at the end of a sentence and the distance with the subordinate words in the next sentence, for example, since “中下 is a term in the surname dictionary and “の”is a subordinate word, “中下旬、本岛の”is wrongly recognized as a name chunk.

Therefore, we build a set of high frequency disturbing characters such as “市”“町”

“館”，alongside a set of punctuations, and remove the candidate which containing elements in the above sets. The test shows that the filtration impurity ratio is over 80%.

In the process of anaphora regression, a temporary Hash table is set to cache the current completed names keyed by each surnames. In this way, when the candidate chunk is an abbreviated or anaphora expression, with the Hash table can we find out the completed name of chunk and make sure the normalization of the results.

3. Results analysis based on graph database

3.1.1. Construction of hot event leader graph based on neo4j graph database

Neo4j is a non relational database, which is flexible, convenient and stable for the management of nodes, attributes and relationships. Based on the natural extensibility of graphs, the continuous tracing of events and the dynamic supplement of characters can be realized.

The hot event leader graph can be described as follows:

1.A root node, which makes a macro description of time axis, and its attributes include the time range and the total reporting frequency of events.

2. 16 Year nodes, which set up according to the corpus time range (2000-2015). The attributes contain the year and the reporting frequency of related events of it.

3. Character node, according to the decreasing frequency of the co-occurrence frequency, select the first 5 characters of the high frequency co-occurrence in each year to create the character node, the attributes include name and its position.

Hot event leader of Yasukuni Shrine during 2000-2015 years can be shown as figure 3.

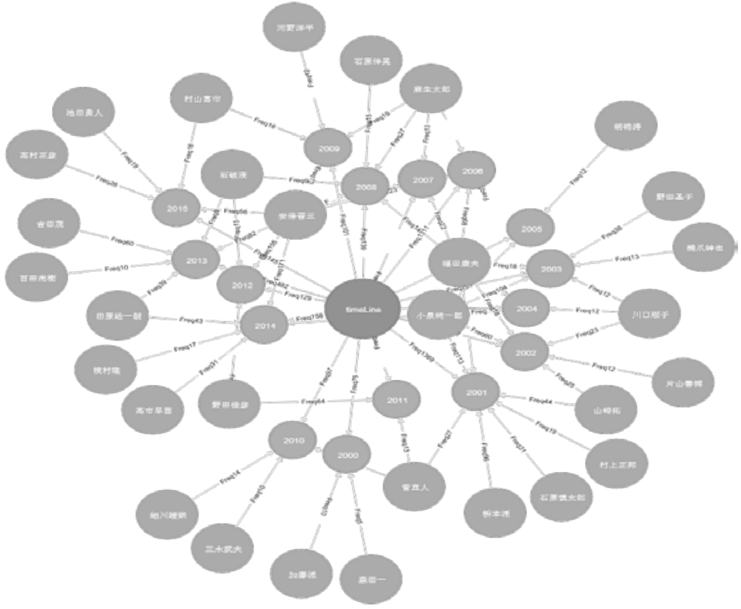

Figure 3 Hot event leader of Yasukuni Shrine from 2000 to 2015.

3.1.2. Evaluation metrics and recognition result

In this paper, 50 test corpora are selected randomly, and the names of them are manually identified. The accuracy, recall rate and F value are used to evaluate the quality of Japanese name recognition. The results are listed as table 1.

Table 1 Statistical table of the test results.

metric	number of the correct recognized names	total number of the recognized names	total number of the names in the corpus	accuracy	recall rate	F value
result	412	454	452	90.07%	91.15%	0.906

3.1.3. Hot events analysis--take Yasukuni Shrine event as an example

The reports number of Yasukuni shrine is a barometer of Sino Japanese relations. The reports number of the Yasukuni Shrine in 2000-2015 is shown in figure 4.

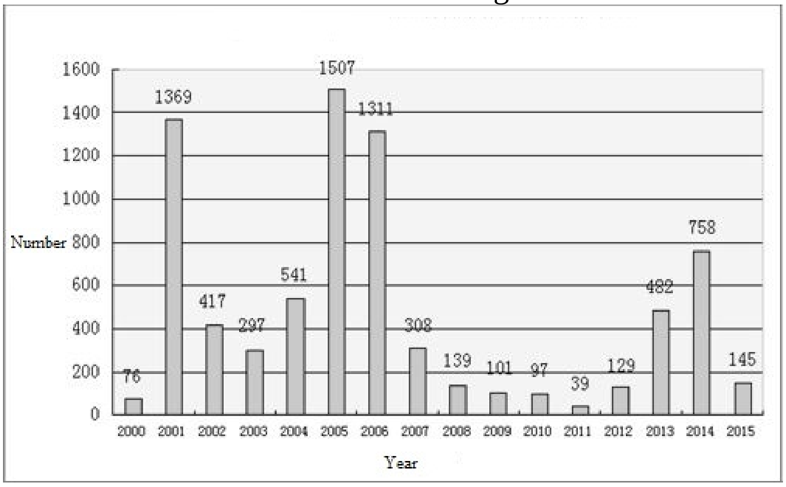

Figure 4 The quantity of report on Yasukuni in Asahi Shimbun in 2000-2015.

When the reports number increased, the relationship between China and Japan was at a low ebb. In 2001, the number increased explosively, as a result of the Junichiro Koizumi administration outrageously Yasukuni Shrine in August 13th. Not alone, in 2005, the number were again blowout, when Sino Japanese relations has hit the lowest temperature since the diplomatic relations establishment for 33 years, because 2005 was the 60th anniversary victory of the world anti fascist war and the Chinese people's Anti Japanese War. On the contrary, when the number declined sharply, the relations tended to ease from the previous year. For example, in 2007, the number of reports turns to a cliff fall and continued to decline in the next 4 years, because of Andouble's first resumption of relations with China after his first appointment. But in October 2012, after Andouble took over the state power again, his attitude and policy towards China became tough, resulted in the highly increased report number in 2014.

High frequency characters of Yasukuni Shrine are the indicators of Japan's attitude towards China. As the core character map of hot events is established through the corpus mining, the attitude of the related figures can be used as clues in the event analysis. From the data point of view, in 2001 Koizumi appeared up to 113 times, even 5 times of the highest frequency in 2000. As a representative of Japanese political hawks, Koizumi took office as Prime Minister of Japan. In 2012, the frequency of Andouble reached 105 times, According to history, in 2012 when Andouble was in power again, his policy towards China changed obviously and his attitude towards the Yasukuni shrine was also more radical.

4. Conclusion

In this paper, we present a hot event analysis method based on Japanese news corpus. The main tasks include corpus construction, name chunk recognition, anaphora regression, neo4j based event leader map modelling and data based hot event analysis. The validity of the method was verified by the relevant corpus of Yasukuni shrine. Results show that our method can fast mining high frequency co-occurrence characters in hot events. By constructing the knowledge map and sorting out the historical materials of core characters, we can grasp the development and macro changes of events more intuitively and clearly.

References

[1]ZHANG Han, (2000) Identification of names in segmentation of Japanese text. (Doctoral dissertation, Dalian University of Technology).

[2]平田亜衣，小町守．様々なジャンルのテキストに対する固有表現認識の分析. In エラー分析ワークショップ(言語処理学会年次大会 2015).

[3]浅原正幸, 松本裕治．日本語固有表現抽出におけるわかち書き問題の解決. 情報処理学会論文誌．2004．

[4]DAI Siming, (2012) Research and application of Internet text hotspot information entity recognition, (Master dissertation, South China University of Technology).

[5]RU Kuang, (2014) Japanese Chinese bilingual named entity to study the acquisition method and its application, (Master dissertation, Beijing Jiaotong University).

[6]YI Cunyan, HUANG Shujian, DAI Xinyu, CHEN Jiajun, (2015) Chinese and Japanese named entity translation extraction for news corpus—Small microcomputer system, 6.

[7]ZHANG Suxiang, ZHANG Suxian, WANG Xiaojie, (2008) A method of person name recognition—Computer engineering and Application, 21.

[8]HE Yanxiang, LUO Chuwei, HU Binyao, (2015) Geographic named entity recognition method based on combination of CRF and rules—Computer application and software, 1.

[9]福島健一，鍜治伸裕，喜連川優．日本語固有表現抽出における超大規模ウェブテキストの利用［J］．情報科学．2012 年 05 期.

[10]福岡健太. Semi-Markov Conditional Random Fields を用いた固有表現抽出に関する研究［J］. https://library.naist.jp/mylime dio/ dllimedio/ showpdf2. cgi/ DLPD FR 003768_P1-75