Detecting Topics Popular in the Recent Past from a Closed Caption TV Corpus as a Categorized Chronicle Data

Hajime Mochizuki[1] and Kohji Shibano[2]
[1]Institute of Global Studies, Tokyo University of Foreign Studies, Tokyo, Japan
[2]Research Institute for Languages and Cultures of Asia and Africa, Tokyo University of Foreign Studies, Tokyo, Japan

Keywords: Topic Detection, Closed Caption TV Corpus.

Abstract: In this paper, we propose a method for extracting topics we were interested in over the course of the past 28 months from a closed-caption TV corpus. Each TV program is assigned one of the following genres: drama, informational or tabloid-style program, music, movie, culture, news, variety, welfare, or sport. We focus on informational/tabloid-style programs, dramas and news in this paper. Using our method, we extracted bigrams that formed part of the signature phrase of a heroine and the name of a hero in a popular drama, as well as recent world, domestic, showbiz, and so on news. Experimental evaluations show that our simple method is as useful as the LDA model for topic detection, and our closed-caption TV corpus has the potential value to act as a rich, categorized chronicle for our culture and social life.

1 INTRODUCTION

Corpora have become the most important resources for studies and applications related to natural language. Various studies and applications of corpus-based computational linguistics, knowledge engineering, and language education have been reported in recent years (Flowerdew, 2011)(Newman et al., 2011). However, the proportion of spoken language in corpora is quite low. There are only a few “spoken language corpora,” such as the Corpus for Spontaneous Japanese (Maekawa et al., 2000) and a part of the British National Corpus (Corpus, 2007), that can be used for research purposes. Under these circumstances, we have been continuing to build a large-scale spoken language corpus from closed-caption TV (CCTV) data. We have been collecting the CCTV corpus since December 2012, and its size has reached over 116,000 TV programs, over 44 million sentences, and 475 million words as of March 2015.

After amassing this spoken-language corpus, we will be positioned to employ it in a wide variety of research areas as a language resource. TV is a major form of media and is familiar in our daily lives. Furthermore, TV broadcasting stations assign each TV program at least one of twelve genre labels, such as “Animation,” “Drama,” and “News.” Therefore, we expect that the CCTV corpus will act as a rich, categorized chronicle for our culture and social life.

In our research, we aimed to extract recent and popular or interesting topics from the CCTV corpus. The variety of information changes rapidly in our modern society, and we often cannot remember something from recent TV episodes or concerning topics of interest to us. For example, few people can quickly describe the prevailing societal issues beyond the most recent two months. It is unexpectedly difficult to remember what dramas, songs, and topics were interested.

In this paper, we propose a simple method for extracting topics that were popular over the past 28 months from the CCTV corpus.

Our research is related to trend or topic detection (Mathioudakis and Koudas, 2010)(Glance et al., 2004)(Weng and Lee, 2011)(Wang et al., 2013)(Mochizuki and Shibano, 2014) (Lau et al., 2012)(Fujimoto et al., 2011)(Keane et al., 2015) and buzzword extraction (Nakajima et al., 2012) studies. Many of these studies aim to extract popular topics or buzzwords from a large number of texts in consumer-generated medias (CGM), such as Weblog articles or tweets, but we analyze closed-caption text from the mass media.

Because of the wide-spread use of the Internet, there are people of the opinion that social media plays a central role in public culture and social movements.

However, we believe that TV programs still have a strong influence on the general public. Despite the

Mochizuki, H. and Shibano, K..
Detecting Topics Popular in the Recent Past from a Closed Caption TV Corpus as a Categorized Chronicle Data.
In *Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2015) - Volume 3: KMIS*, pages 342-349
ISBN: 978-989-758-158-8

cultural contribution people make with Twitter messages about specific topics, the spread of that topic to the many others who do not use Twitter will be limited. On the other hand, if such a topic is reported by TV programs even once, the topic will be well known to the general public, including those who do not usually use the Internet. Topics thought to be valuable to the public will be repeatedly reported in many TV programs. Words related to the topic will therefore occur frequently in the voice data of these TV programs.

In this paper, we focus on three genres: information/tabloid-style programs, referred to as "genre I"; drama, referred to as "genre D,"; and news, referred to as "genre N." Among these three genres, we center genre I as a nuclear genre, because information/tabloid-style programs consist of miscellaneous topics that are also included in genres N or D. We expect that a word that occurred frequently in genre I and which was treated by many TV stations has a high possibility of being a popular topic that captured the attention of the public. After detecting a frequent word in genre I, we decide on the sort of topic by investigating the appearance tendency of the same word in genres N and D.

To evaluate the effectiveness of our simple method, we extracted topics by Latent Dirichlet Allocation (LDA), which is one of the most popular methods for topic detection(Blei et al., 2003), and compared both methods' topics. From these results, we will show that our simple method is as useful as LDA, and that our TV closed-caption corpus is useful material that includes various kinds of popular topics.

Unfortunately, it would be impossible to release the CCTV corpus itself because of a legal issue. As an alternative method, we should mention how to build a corpus from TV data for researchers who wish to use CCTV data. Therefore, we will describe the details of the CCTV corpus in the next section.

2 COLLECTING CLOSED CAPTION TV DATA

2.1 Japanese Television Services with Closed Caption

In Tokyo, there are seven major broadcasting stations that organize the nationwide Japanese network: (1) NHK-G, (2) NHK-E, (3) NTV, (4) TBS, (5) TV Asahi, (6) Fuji TV, and (7) TV Tokyo. One of the characteristics of Japanese TV stations is that they provide a wide variety of programs belonging to different genres. According to the classifications provided by the EPG (Electric Program Guide), there are at least 12 genres: "Animation," "Sport," "Culture and Documentary," "Drama," "Variety," "Film," "Music," "Hobby and Educational," "Inform Tabloid-Style," "Welfare," and "Other." According to a report by the Ministry of Internal Affairs and Communications detailing the achievements of closed-caption TV in 2012[1], the amount of programming that included closed-caption data reached approximately 50%.Therefore, a large number of resources for building a spoken language corpus are currently available in Japan.

We can use definition of (ARIB, 2009) to extract closed-caption data. The following three procedures are necessary for building a CCTV corpus: (1) Record all TV programs with closed caption data in a TS (Transport Stream) format during a 24-hour period; (2) automatically extract the closed-caption data in ASS (Advanced SubStation Alpha) format from the TS data; (3) filter the ASS-format file to extract a plain text format file and execute a morphological analyzer; and (4) convert the video data in TS format into MP4 format. Some open-source software applications are available for these purposes.

2.2 Recording All TV Programs Television Services with Closed Caption

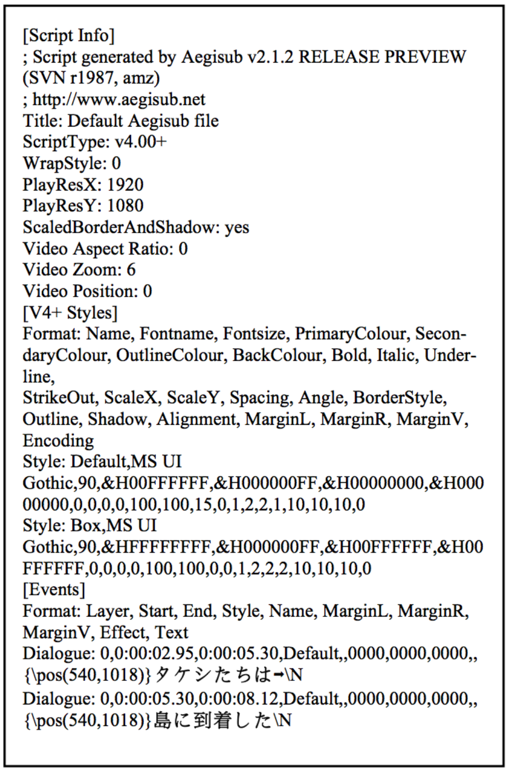
We used the freeware packages *EpgDataCap_Bon* and *EpgTimer* to record all TV programs in Tokyo.*EpgTimer* can be used to retrieve the EPG (Electronic Program Guide) list and set the timer to record. *EpgDataCap_Bon* is executed by *EpgTimer* to record the programs, generating a TS-format file and a program information file for each program. The TS-format file is a full segment broadcasting video file. The program information file includes certain information, such as the program name, broadcast time, station name, and a program description.

2.3 Extracting Closed Caption Data

The next procedure is to extract the closed caption data in ASS format from the TS-format file. The closed caption data are mixed with video data and transmitted through digital terrestrial broadcasting. Therefore, we must use a special program to separate the closed caption data from the TS format file. *Caption2Ass_PCR*, a freeware program, is available

[1] http://www.soumu.go.jp/menu_news/s-news/01ryutsu09_02000071.html

for this purpose. An example of an ASS format file is shown in Figure 1 and an example of a screen image of closed-caption TV in Figure 2.



```
[Script Info]
; Script generated by Aegisub v2.1.2 RELEASE PREVIEW
(SVN r1987, amz)
; http://www.aegisub.net
Title: Default Aegisub file
ScriptType: v4.00+
WrapStyle: 0
PlayResX: 1920
PlayResY: 1080
ScaledBorderAndShadow: yes
Video Aspect Ratio: 0
Video Zoom: 6
Video Position: 0
[V4+ Styles]
Format: Name, Fontname, Fontsize, PrimaryColour, Secon-
daryColour, OutlineColour, BackColour, Bold, Italic, Under-
line,
StrikeOut, ScaleX, ScaleY, Spacing, Angle, BorderStyle,
Outline, Shadow, Alignment, MarginL, MarginR, MarginV,
Encoding
Style: Default,MS UI
Gothic,90,&H00FFFFFF,&H000000FF,&H00000000,&H000
00000,0,0,0,0,100,100,15,0,1,2,2,1,10,10,10,0
Style: Box,MS UI
Gothic,90,&HFFFFFFFF,&H000000FF,&H00FFFFFF,&H00
FFFFFF,0,0,0,0,100,100,0,0,1,2,2,2,10,10,10,0
[Events]
Format: Layer, Start, End, Style, Name, MarginL, MarginR,
MarginV, Effect, Text
Dialogue: 0,0:00:02.95,0:00:05.30,Default,,0000,0000,0000,,
{\pos(540,1018)}タケシたちは➡\N
Dialogue: 0,0:00:05.30,0:00:08.12,Default,,0000,0000,0000,,
{\pos(540,1018)}島に到着した\N
```

Figure 1: Example of an ASS format file.

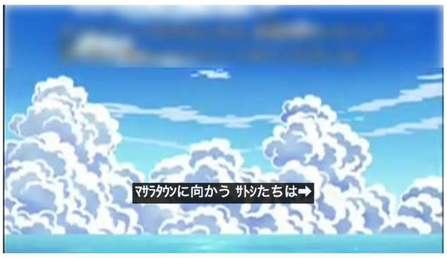
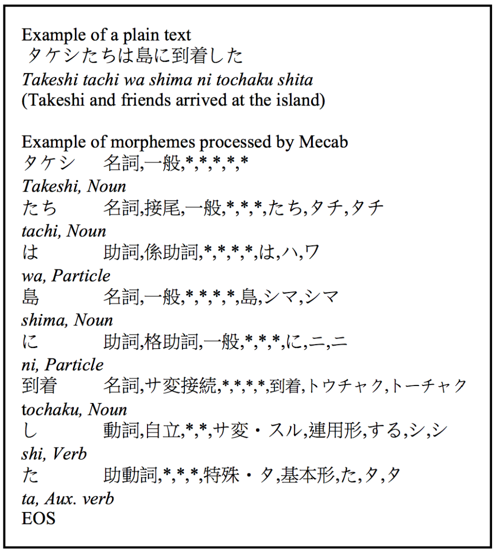


Figure 2: Example of a screen image of closed-caption TV data. This example is a screen shot from the animation program “the Pocket Monsters” broadcasted in Tokyo on May 2, 2013.

As shown in Figure 1, the ASS file consists of at least three parts: “[Script Info],” “[V4+ Styles],” and “[Events].” The first, the ”[Script Info]” section, contains information about the closed-caption file, such as its title, creator, type of script, and display resolution. In this example, the second line shows that the text file was generated by Aegisub, open-source software used to create closed-caption data. The eighth and ninth lines show that the display resolution is 1920x1080 pixels. The second section, “[V4+ Styles],” provides a list of style definitions, such as font, font size, and primary/secondary/outline/background color. In this study, we use the style “Name” as a signifier for closed-caption texts. We only use closed-caption texts in which the style is classified as “Default.” We ignore the “Rubi” style because “Rubi” indicates Japanese kana, which are used to show the pronunciation of Kanji characters. The third section, “[Events]” is the list of closed-caption text that the TV station intends to display with a particular timing. The creator of closed caption text can specify certain attributes, such as the style to use for this “event,” the position of the text on the TV screen, display timing, and text to display. Information about timing is divided into start time and end time, and is formatted as h:mm:ss. The bottom two lines in Figure 1 are examples of closed-caption texts. In these examples, we use the two Japanese strings in these lines as the content of our corpus.

2.4 Filtering ASS file to Generate Plain Text and Create Morphemes Data

For post-processing, we filtered the ASS-format files to generate plain language text files without meta-symbols. These plain texts are composed of Japanese sentences, and are finally divided into morphemes tagged with the part-of-speech information. We used MeCab (Kudo et al., 2004) as the morphological analyzer. Examples of a plain text and morphemes-format file are shown in Figure 3.



```
Example of a plain text
 タケシたちは島に到着した
Takeshi tachi wa shima ni tochaku shita
(Takeshi and friends arrived at the island)

Example of morphemes processed by Mecab
タケシ     名詞,一般,*,*,*,*,*
Takeshi, Noun
たち       名詞,接尾,一般,*,*,*,たち,タチ,タチ
tachi, Noun
は         助詞,係助詞,*,*,*,*,は,ハ,ワ
wa, Particle
島         名詞,一般,*,*,*,*,島,シマ,シマ
shima, Noun
に         助詞,格助詞,一般,*,*,*,に,ニ,ニ
ni, Particle
到着       名詞,サ変接続,*,*,*,*,到着,トウチャク,トーチャク
tochaku, Noun
し         動詞,自立,*,*,サ変・スル,連用形,する,シ,シ
shi, Verb
た         助動詞,*,*,*,特殊・タ,基本形,た,タ,タ
ta, Aux. verb
EOS
```

Figure 3: An example of plain text extracted from an ASS file and its morpheme data processed using MeCab.

2.5 Creating the Closed Caption TV Corpus

We have currently recorded a total of 116,583 programs, or 73,832 hours and 13 minutes of programming. Each program is classified into at least one genre.

Table 1 shows the total number of morphemes by genre. The scale of our corpus at this point is 475,096,596 morphemes from 44,951,572 sentences. We can state that our corpus is presently one of the largest spoken Japanese corpora. In this study, we selected and used the sub-corpora from three genres: drama, labeled D; news, labeled N; and information and tabloid-style programs, labeled I. The sizes of the sub-corpora are 12,061 programs, 22,043 programs, and 16,624 programs, respectively. Approximately 43% of our corpus is made up of 50,728 total programs.

3 METHOD FOR DETECTING RECENT TOPICS

Our purpose in this paper is to detect the topics and episodes of recent interest to us from a large-scale TV closed-caption corpus. We focus on words related to genres I, N, and D. We start with the words treated in genre I and investigate the appearance tendency of the same words in genres N and D. When a drama is especially successful, topics related to it also tend to become popular and to appear in genres I and D. The words that appear with the same tendency in both genres I and N tend to express topics related to news articles. We estimate that this phenomenon can be observed by comparing the word distribution among the three genres.

For example, attractive heroes or main characters in a hit drama would also become popular. Words related to the characters, such as their names, fashions, and signature dialogue, may frequently occur not only in dramas (genre D) but also in the information/tabloid-style genre (genre I).

On the other hand, words related to an unpopular drama might not appear in genres other than dramas if the drama went unnoticed by other TV programs. We adopted the following process to discover topics of past interest to us. Besides, we use bigrams, which are pairs of adjacent two words, instead of single words, in this paper. We think bigrams are more suitable for forming topics than single words, because single words tended to be too small as topic words in our previous work.

(1) Counting the total number of bigrams and the number of each word every month. We express the monthly distributions for each bigram by its monthly proportion.

(2) Picking out bigrams from genre I that appear in specific months. These bigrams have the possibility of being the characteristic bigrams in specific topics.

(3) Investigating how the bigrams selected in Step (2) appear in other genres (genre N and D). It is expected that if one bigram is related to a popular drama, the distribution of the bigram in genre D is similar to the distribution in the genre I. Similarly, it is also expected that if one bigram is related to an interesting topic, such as a significant event or crime, the distribution of the bigram in genre N is similar to the distribution in genre I.

(4) Checking whether the selected bigrams from Step (3) can be considered to express the topics of interest to us in the past. During this step, we must confirm the answer by checking texts that contain the target bigrams in the corpus for the present.

In Step (1), the total occurrence number of bigram i is calculated with Equation 1.

$$TotalFreq_i = \sum freq_{i,j} \tag{1}$$

where $freq_{i,j}$ refers to the frequency of the bigram i in month j. The $ratio_{i,j}$ that is the proportion of bigram i in month j, is calculated with Equations 2 and 3.

$$biasedFreq_{i,j} = freq_{i,j} \times \frac{N}{N_j} \tag{2}$$

$$ratio_{i,j} = \frac{biasedFreq_{i,j}}{\sum_n biasedFreq_{i,n}} \tag{3}$$

where N is the total number of bigrams in the sub-corpus, and N_j is the total number of bigrams in the j-th month in the sub-corpus. The j varies from 1 to 28 because our corpus has terms of 28 months. Therefore, each variable of frequency ratio for each bigram in a sub-corpus has 28 values.

In step (2), we set three thresholds: the bigram frequency, the monthly frequency, and the frequency ratio per month.

For the first threshold, we select bigrams that have a frequency of more than one per million bigrams.

For the second threshold, we use a metric similar to the document frequency (DF) in the information retrieval domain. The DF for term i refers to the number of documents that contain the term i. The DF is considered to represent the level of a term's specificity in the entire document set. We use a monthly frequency, referred to as MF, instead of the DF. If bigram i appears in all months of our corpus, the MF

Table 1: Total number of morphemes for each genre.

Genre	No. of Progs.	Total Hrs (h:m)	No. of Sents.	No. of Morphes
A: Animation	9,959	3,612:49	2,961,139	22,053,412
S: Sport	4,313	3,609:09	1,991,883	24,087,726
C: Culture/Documentary	14,313	6,411:42	3,236,149	38,479,292
D: Drama	12,061	10,675:25	7,698,865	66,055,199
N: News	22,043	15,056:48	6,087,279	109,170,849
V: Variety	19,462	17,201:08	12,687,446	107,019,473
F: Film	711	1,413:04	867,138	6,373,679
M: Music	1,927	1,283:59	496,616	5,374,042
H: Hobby/Educational	12,930	3,956:44	3,057,684	27,480,760
I: Information/Tabloid Style	16,624	9,621:34	5,218,918	61,728,141
W: Welfare	1635	693:56	374,138	4,553,119
O: Other	605	300:16	274,317	2,720,904
Total	116,583	73,836:34	44,951,572	475,096,596

for bigram *i* is 28, because our corpus was collected over 28 months. We seek bigrams with monthly frequencies of less than 18 months.

For the third threshold, we seek biased bigrams with a monthly frequency proportion of at least 10%. This 10% threshold means that the frequency of the bigram is biased by approximately twice that of the average proportion, because the average frequency proportion for 18 months is approximately 5.5%. We think that it is natural that when a topic captured public attention, words related to the topic will occur frequently. We extract the bigrams that satisfy all of these conditions.

In Step (3), we consider only the bigrams selected in Step (2). For the other genres, we use the sub-corpora from genres D and N. We check whether the distribution for a selected bigram in Step (2) is similar to its distributions in both genres D and N. We remove the bigrams that have fewer similarities. We can set the similarity criteria. In this study, we use the following criteria to select bigram *i*.

(3-1) The monthly frequency of bigram *i* from the genre I sub-corpus is the same as or larger than in genres D or N.

(3-2) At least one of the monthly frequency proportions is over 10% in the genre I sub-corpus.

(3-3) Pearson's correlation coefficient *r* between genres I and D, or between genres I and N, equals 0.7 or higher. We calculate and check *r* for each bigram.

During Step (4), we check whether the bigrams selected during Step (3) can be considered to express recent popular topics. We expect that the bigrams related to a hit drama or various kinds of news can be detected during these steps.

4 EXPERIMENTAL EVALUATION

To determine the effectiveness of our method for detecting past topics of interest, we performed an experiment and examined how well it performed with bigrams.

4.1 Counting the Bigram Frequency

We counted the frequency of bigrams in our corpus prior to the experiments.

The numbers of bigram types in genres I, D, and N are 3,158,472, 3,101,213, and 3,541,631, respectively. The numbers of bigram tokens are 65,256,640, 71,908,595, and 112,222,068, respectively. Among them, the numbers of types that have a frequency of more than one per million are 83,748 in genre I, 77917 in genre D, and 88,484 in genre N. These ratios are 2.7%, 2.5%, and 2.5%, respectively. Similarly, the numbers of types that have a frequency of more than one per million in genres I, D, and N are 52,077,048, 57,784,431, and 92,761,338, respectively.These ratios are 79.8%, 80.4%, and 82.7%, respectively.

As shown in Table 3, the total frequency for the top 2.5% or 2.7% of bigrams is about 80%. Approximately 2.5% or 2.7% of all three genres' vocabularies have a coverage of 80% in the sub-corpora.

4.2 Experiment Involving Extracting for Bigrams between Genres I and D

The first experiment we report assessed the topics that were mined with our method using the frequency of bigrams in genres I and D. First, we extracted candidate bigrams from the results in Step (2), as described in Section 3.

Table 2: The extracted bigrams and dramas.

Drama (bigrams): Example(Pearson's *r*), ...
Ama-chan (13): *Aki/chan*(0.82), *j/eje*(0.86), *kita/sanriku*(0.87), *no/Ama*(0.84)
Massan (11): *Massan/no*(0.94), *Ellie/san*(0.91), *Whisky/wo*(0.89), *Ellie/chan*(0.91)
Gochiso-san (7): *Me/iko*(0.84), *Yuta/ro*(0.87), *Kazue/san*(0.72)
Hanako-to-Ann (6): *Hana/chan*(0.82), *Renko/san*(0.73), *Daigo/san*(0.80)

Table 3: The extracted bigrams and news.

Category (bigrams): Example(Pearson's *r*), ...
World (67): Islamic/State(0.92), Snowden/*shi*(0.98), Ebola-hemorrhagic-fever/*no*(0.93), Pro/Russian(0.95), Korean-air-lines/*no*, Jordan/government(0.95)
Domestic (85): unsafe/drug(0.82), d/rone(0.95), bird/flu(0.98), *Hokuriku*/super-express(0.94), *Obuchi*/*san*(0.85), *Kako/sama*(Princess Kako)(0.80)
Showbiz (31): *Tsunku*/*san*(0.93), ghost/writer(0.98), *Takakura*/*san*(0.99)
Sport (51): *Hanyu*/*senshu*(0.94), a triple/axel(0.97), *Nishikori*/*senshu*(0.92), *Hakone*/*ekiden*(0.87), Zacch/Japan(0.87), World-Cup/*ni*(0.73)
Weather (64): strong/typhoon (0.96), *bofu/setsu*(blizzard)(0.77), hot/summer(0.91)
Culture (28): *obake/yashiki*(a haunted house) (0.83), hallo/ween(0.90), *nengajo*(new year card)/*wo* (0.77), Xmas/present(0.93),*Shikinen/Sengu*(0.82)
Economy (10): *Kumiko*/*shacho* (chairwoman)(0.99), *Ootsuka-Kagu* (furniture shop)/*no*(0.99), oil/price(0.76), *kabunushi/sokai* (stockholders' meeting)(0.90)
Science (9): STAP/cell (0.87), (*Obo*)*kata/Haruko*(0.98), Professor/*Amano*(0.98)

For this experiment, first we extracted 3,507 bigrams from genre I as candidate bigrams related to popular dramas. Next, we counted the frequency of the same 3,507 bigrams in genre D and selected 422 bigrams that had frequencies of more than one per million. In Step (3), we investigated how these 422 bigrams appeared in genres I and D. We found bigrams that had similar distributions in genre pairs I and D. In fact, we calculated and checked Pearson's correlation coefficient *r* for each bigram between genres I and D. From the result in Step (3), we found 56 bigrams which their *r* equal 0.7 or higher.

In Step (4), the final step, we checked whether the bigrams selected during Step (3) could be considered related to recently popular dramas. Out of 56 bigrams, 41 were related to dramas. We could not find relations to specific dramas for 15 bigrams.

The results show that thirteen, eleven, seven and six bigrams were related to the great hit dramas, *Ama-chan*, *Massan*, *Gochiso-san* and *Hanako-to-Ann*, respectively. The remaining four bigrams were related to three different dramas that were less popular. Table 2 shows the titles of the four dramas and examples of bigrams.

From a different viewpoint, 29 bigrams were names or nicknames of characters in dramas. Five bigrams were related to key items of dramas such as *whisky* and *shiosai*. Two bigrams were the titles of dramas. Only three bigrams were a part of a signature phrase of a heroine, and only one bigram was the name of a location in a drama.

In Table 2, *Ama-chan* was the greatest hit drama in 2013, and *Aki-chan* is the heroine of *Ama-chan*. In Japan, everybody knows her. *Jejeje* is a signature phrase of hers, and was also the winner of the Keywords-of-the-Year contest for 2013. It can be said that our method in the first experiment tended to yield parts of character names or signature phrases in past dramas, including greatly popular dramas. However, the effectiveness of our method was limited to only four dramas that were broadcasted by NHK-G from Monday to Saturday over six months.

4.3 Experiment Involving Extracting for Bigrams between Genres I and N

The second experiment we report assessed what topics were mined with our method using bigrams in genres I and N. Like Experiment 1, first we extracted 3,507 bigrams from genre I as candidate bigrams related to news topics.

Next, we counted the frequency of the same 3,507 bigrams in genre N and selected 911 bigrams that had frequencies of more than one per million. In Step (3), we investigated how these 911 bigrams appeared in genres I and N. In fact, we calculated and checked Pearson's correlation coefficient *r* for each bigram between genres I and N. From the results of Step (3), we found 454 bigrams whose *r* equaled 0.7 or higher.

In Step (4), the final step, we checked whether the bigrams selected during Step (3) could be considered related to recent interesting news. Out of 454 bigrams, 345 could be found with relations to any news topic. We could not determine the relation to specific news for 109 bigrams because their meanings were too general.

Table 3 shows eight categories of news topics and examples of bigrams. The results show that sixty-seven, eighty-five, thirty-one, fifty-one, sixty-four, twenty-eight, ten, and nine bigrams were related to world, domestic, showbiz, sports, weather, culture, economy and science news topics, respectively.

It seems that many extracted bigrams reflect various aspects of the present society. Unfortunately, many of them were not good news. Many topics from the world news were related to terrorism, the Islamic State, and Edward Snowden. The reason why there was a large number of topics related to the weather reflects recent extraordinary weather throughout the world. Many topics from science news were related to the STAP cells scandal. However, we extracted some pleasurable topics, such as the opening of the *Hokuriku Shin-kansen*, the new Japanese railroad, and excellent athletic performances by *Yuzuru Hanyu* and *Kei Nishikori*.

The results of these two experiments are not inconsistent with our memories and sense impressions. It can be said that our method of using genres I and N can yield recently popular news topics.

Table 4: The topics extracted by LDA.

score	words
genre I	
0.021	America, president, North Korea, Russia, Obama, Ukraine, ...
0.007	Isramic, Jordan, Syria, Iraq, Turkey, Japanese
0.030	Live, Debut, concert, AKB, Fun, dance, Idol
0.022	Mt. *Ontake-san*, East Japan, earthquake
0.009	Virus, influenza, allergy, vaccine, asthma, a sideeffect,...
0.009	*Massan*, *Ellie*, Wisky, ...
0.022	Drama, *Kanbei*, scene, *Taiga*, Hero, ...
genre D	
0.014	*Massan*, *Ellie*, Wisky, Scotland,...
0.160	date, party, girls,...
0.103	policemen, PC, police, net, stalker,...
0.023	*Kanbei*, *Hanbei*, *onago*, *gozaru*, child,...
0.018	Idol, gege, GMT, memory, mother, cafe,...
genre N	
0.199	truck, bicyle, car, police, taxi, intersection, driver,...
0.096	Obama, president, America, Washington, Snowden,...
0.078	*Tokyo Denryoku*, atomic energy, tank, radioactive rays,...
0.075	goal, soccer, team, league, World cup,...
0.055	Russia, Ukraine, America, President, Europe
0.037	Virus, influenza, Ebola, vaccine, WHO,...
0.018	Islamic, Jordan, Japanese, pilot, jornalist,

4.4 Supplementary Experiment Using LDA

This research is potentially related to event or topic detection. Latent Dirichlet Allocation (LDA) is a well-known and popular topic model for event or topic detection(Blei et al., 2003) . It is a powerful model for analyzing massive sets of text data. Many works on topic detection have used LDA (Lau et al., 2012)(Fujimoto et al., 2011)(Keane et al., 2015).

Here, we also tried to apply LDA to our event detection. We tried to detect topics in text sets of genres I, N, and D by the LDA of the Mallet language toolkit (McCallum, 2002). We set the number of topics to be used at 100, and the other parameters used the Mallet defaults. Part of the extracted words are shown in Table 4. In the Table, “score” means Dirichlet coefficient.

Because this method of extracting topics is different from our method, it cannot compare the results directly. We investigated similar word groups between genres I and D, and genres I and N. The results between genres I and D show that there are two similar word groups, related to the TV series *Massan* and *Kanbei.* The results between genres I and N show that there are three similar word groups, related to *Isramic, Ebola,* and *Russia and Ukraine.*

Relatively few word groups related to two genres were detected in this supplementary experiment using LDA.

5 DISCUSSION

In the first two experiments, there was no overlap in the 1,333 bigrams made up of the 422 bigrams in genre D and the 911 bigrams in genre N. Therefore, the total number of candidate bigrams was 1,333 out of 3,507, and we extracted 510 bigrams whose Pearsons r in genre I equaled 0.7 or higher. Finally, 395 out of 510 bigrams that had relations to topics from specific dramas or news articles were found.

All four dramas related to the extracted bigrams in genre D were big hits from the recent 28 months. Many extracted bigrams from genre N were related to topics that captured public attention during these 28 months.

These results have encouraged us to use our method of determining the sort of topic after detecting a frequent word in genre I by investigating the appearance tendency of the same word in genres N and D. The results also support our expectation that the CCTV corpus could act as a rich categorized chronicle data.

On the other hand, it seems that a method based on high frequency is too strict for genre D. Since there are a few hit dramas that were unable to be detected by our method, the threshold of frequency should be lowered in genre D.

In the third supplementary experiment, relatively few word groups related to two genres were detected. Though our method is not a more sophisticated model than LDA, the results seem to show that suitable topics can be detected with our method.

6 CONCLUSION

In this paper, we described methods for detecting past topics of interest in a CCTV corpus. The results show that some bigrams related to hit dramas and popular or interesting news were extracted.

Experimental evaluations show that our simple method is as useful as the LDA model for topic detection, and our CCTV corpus has the potential value to act as a rich, categorized chronicle for our culture and social life.

To improve the accuracy and availability of our method, we intend to pursue the following future projects to further our work in detecting recent topics: (1) Reduce the threshold of frequency in genre D when the similarity between I and D is calculated; (2) Compare genre I with genres other than D and N; (3) Count the total number of bigrams and the number of each bigram every week, rather than every month; (4) Use a longer unit instead of a bigram.

ACKNOWLEDGEMENTS

This research was supported by JSPS KAKENHI Grant Numbers 26240051 and 15H02794.

REFERENCES

ARIB (2009). Service information for digital broadcasting system (in japanese). In *Association of Radio Industries and Businesses*.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *The Jornal of Machine Learning Research archive*, 3.

Corpus, B. N. (2007). The British National Corpus, version 3 (BNC XML Edition). *Oxford University Computing Services on behalf of the BNC Consortium*. http://www.natcorp.ox.ac.uk.

Flowerdew, L. (2011). *Corpora and Language Education*. Palgrave Macmillan.

Fujimoto, H., Etoh, M., Kinno, A., and Akinaga, Y. (2011). Topic analysis of web user behavior using lda model on proxy logs. *Advances in Knowledge Discovery and Data Mining, LNCS*, 6634/2011:525–536.

Glance, N. S., Hurst, M., and Tomokiyo, T. (2004). BlogPulse: Automated Trend Discovery for Weblogs. In *Proceedings of WWW2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*.

Keane, N., Yee, C., and Zhou, L. (2015). Using topic modeling and similarity thresholds to detect events. In *the 3rd Workshop on EVENTS at the NAACL-HLT 2015*, pages 34–42.

Kudo, T., Yamamoto, K., and Matsumoto, Y. (2004). Applying conditional random fields to japanese morphological analysis. In *EMNLP 2004, the Conference on Empirical Methods in Natural Language Processing*, pages 230–237.

Lau, J. H., Collier, N., and Baldwin, T. (2012). On-line trend analysis with topic models: #twitter trends detection topic model online. In *COLING 2012*, pages 1519–1534.

Maekawa, K., Koiso, H., Furui, S., and Isahara, H. (2000). Spontaneous Speech Corpus of Japanese. In *Proceedings of the Second International Conference on Language Resources and Evaluation LREC2000*, pages 947–952.

Mathioudakis, M. and Koudas, N. (2010). TwitterMonitor: Trend Detection over the Twitter Stream. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2010*, pages 1155–1158.

McCallum, A. K. (2002). Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

Mochizuki, H. and Shibano, K. (2014). Building very large corpus containing useful rich materials for language learning from closed caption tv. *E-Learn 2014, World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education*, 2014(1):1381–1389.

Nakajima, S., Zhang, J., Inagaki, Y., and Nakamoto, R. (2012). Early detection of buzzwords based on large-scale time-series analysis of blog entries. In *ACM Hypertext 2012, 23rd ACM Conference on Hypertext and Social Media*, pages 275–284.

Newman, H., Baayen, H., and Rice, S. (2011). *Corpus-based Studies in Language Use, Language Learning, and Language Documentation*. Rodopi Press.

Wang, J., Zhao, X. W., Wei, H., Yan, H., and Li, X. (2013). Mining New Business Opportunities: Identifying Trend related Products by Leveraging Commercial Intents from Microblogs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1337–1347.

Weng, J. and Lee, B. S. (2011). Event detection in twitter. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pages 401–408.