Re-mining Topics Popular in the Recent Past from a Large-Scale Closed Caption TV Corpus

Hajime Mochizuki and Kohji Shibano

***Abstract*—In this paper, we propose a method for extracting topics we were interested in over the course of the past 18 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, and sport. We focus on dramas and informational/tabloid style programs in this paper. As the results, we extracted some words or bigrams that formed part of a signature phrase of a heroine and the name of a hero in a popular drama.**

***Index Terms*—Topic detection, spoken language corpus, closed caption TV data, word frequency, Pearson's r.**

I. Introduction

Corpora have become the most important resources for researches and applications related to natural language, and a variety of researches and applications for corpus-based computational linguistics, knowledge engineering, and language education have been reported in recent years [1], [2]. Corpora are becoming larger with the increase in machine-readable language resources such as Web pages, wired newspapers, and social media.

Almost all existing corpora are "written language corpora," and only a few "spoken language corpora" such as the Corpus for Spontaneous Japanese (CSJ) [3] can be used for research purposes. To make a spoken language corpus, it is generally necessary to record and dictate voice data. Therefore, a significant amount of time and effort is required to collect and maintain a spoken language corpus as compared to a written corpus, which can be directly collected from Web pages, newspaper articles, and other written materials. Spoken language is used to keep communication in the main part of our intelligent activities. In the fields of computational linguistics, social science, and language education, there is a large significance for spoken corpora as the fundamental data type, and collections of spoken language corpora are currently in large demand.

For our project, we are constructing a large-scale spoken language corpus from closed caption data transmitted through digital terrestrial broadcasting [4]. Over 70% of the recent programs in Japanese digital terrestrial broadcasting use closed caption data, which has been promoted by the Japanese government. Our preliminary investigation has shown that the closed caption scripts for these programs are nearly the same as the words actually spoken in the TV programs. We therefore believe that a very large spoken language corpus can be constructed by collecting closed caption data on a daily basis. We collected the closed caption data from over 70,000 TV programs from January 2013 to June 2014. The total number of words in our corpus has reached over 280 million morphemes.

After amassing this spoken-language corpus, we will be positioned to employ it in a wide variety of research areas as a language resource. Because TV is a major media and familiar in our daily lives, we expect to apply the corpus to language education, offering realistic examples of conversation in e-learning systems.

Elsewhere, TV is a medium for information on culture, sport, and current events. Thus, we expect that keywords related to recent and popular events can be extracted from the corpus such that it will act as a rich chronicle for our culture.

In this paper, we describe a method for extracting recent and popular topics from our TV closed-caption corpus. The proposed method is useful because the variety of information changes rapidly in our modern society, and we often experience something we cannot remember from recent TV episodes or concerning topics we were interested in. For example, there are few people who can quickly describe the prevailing societal issues beyond two months in the past. It is difficult to remember what dramas were popular, what songs were popular and what topics we were interested in.

Because of the wide spread 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 cultural contribution people make with Twitter messages concerning specific topics, the spread of that topic to the many others who do not use the Twitter will be limited.

On the other hand, if such a topic is reported by TV news 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 reported in many TV programs repeatedly. Words related to the topic will therefore occur frequently in the voice data of these TV programs.

In this paper, we propose a method for extracting topics that were popular over the past 18 months from a TV closed-caption corpus. Each TV program is assigned one of the following genres: drama, informational or general

Manuscript received September 10, 2014; revised January 11, 2015. This research was supported by the Grant-in-Aid for Scientific Research (A) (No. 26240051) of JSPS.

Hajime Mochizuki is with the Institute of Global Studies, Tokyo University of Foreign Studies, 3-11-1 Asahi, Fuchu, Tokyo, 183-8534, Japan (e-mail: motizuki@tufs.ac.jp).

Kohji Shibano is with the Research Institute for Languages and Cultures of Asia and Africa, Tokyo University of Foreign Studies, Japan (e-mail: shibano@aa.tufs.ac.jp).

DOI: 10.7763/IJFCC.2015.V4.364

program, music, movies, culture, news, variety, welfare, and sport. We focus on dramas and informational/tabloid style programs in this paper. The size of the sub-corpora for dramas and information/tabloid style shows are 42 million morphemes and 73 million morphemes, respectively.

Our research will be related to a trend detection [5]-[7] or buzzword extraction [8] researches. Many of these researches aim to extract popular topics or buzzwords from a large amount of texts in the consumer-generated medias (CGM) such as Weblog articles or Twitter tweets, though we analyze closed caption text from the mass media.

II. Collecting Closed Caption TV Data

In this section, we describe our corpus collected from closed caption data transmitted through digital terrestrial broadcasting.

A. Japanese Television Services with Closed Caption

In Tokyo, there are seven major broadcasting stations which organize a Japanese nationwide.

1) NHK-G: Japan Broadcasting Corporation.
2) NHK-E: Japan Broadcasting Corporation, with an emphasis on educational programs.
3) NTV: Nippon Television.
4) TBS: Tokyo Broadcasting System.
5) TV Asahi.
6) Fuji TV: Fuji Television.
7) TV Tokyo.

The amount of programming that includes closed caption data has reached about 70%. Therefore, a large amount of resources for building a spoken language corpus is currently available in Japan. The Japanese standard for a closed caption TV system is defined in [9]. We can use this definition to extract closed caption data. The following three procedures are necessary for building our corpus.

1) Record all TV programs with closed caption data in a "TS (Transport Stream) format" during a 24 h period.
2) Automatically extract the closed caption data in "ASS (Advanced SubStation Alpha) format" from TS data.
3) Filter the ASS format file to extract a plain text format file, and execute a morphological analyzer.
4) Convert video data in TS format into MP4 format.
5) Some open-source software applications are available for these purposes.

B. Recording All TV Programs

We use 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.

C. 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 for this purpose. An example of an ASS format file is shown in Fig. 1. The ASS file includes meta-symbols such as information regarding the display location, display duration, and string color, and so on.

```
[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, SecondaryColour, OutlineColour, BackColour, Bold, Italic, Underline,
StrikeOut, ScaleX, ScaleY, Spacing, Angle, BorderStyle, Outline, Shadow, Alignment, MarginL, MarginR, MarginV, Encoding
Style: Default,MS UI Gothic,90,&H00FFFFFF,&H000000FF,&H00000000,&H00000000,0,0,0,0,100,100,15,0,1,2,2,1,10,10,10,0
Style: Box,MS UI Gothic,90,&HFFFFFFFF,&H000000FF,&H00FFFFFF,&H00FFFFFF,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
```

Fig. 1. Example of an ASS format file.

TABLE I: The Numbers and Hours of TV Programs

Date	No. of programs	Total hours
2013.01	2,072	1,269h 58m
2013.02	2,057	1,400h 17m
2013.03	2,173	1,577h 47m
2013.04	3,889	2,366h 45m
2013.05	4,404	2,585h 58m
2013.06	4,487	2,702h 02m
2013.07	4,104	2,550h 57m
2013.08	3,555	2,283h 47m
2013.09	4,224	2,601h 36m
2013.10	4,302	2,689h 35m
2013.11	4,367	2,702h 30m
2013.12	4,247	2,805h 32m
2014.01	4,237	2,847h 19m
2014.02	3,934	2,551h 53m
2014.03	3,925	2,507h 11m
2014.04	4,483	2,832h 30m
2014.05	4,848	3,005h 14m
2014.06	4,729	3,028h 48m
Total	70,037	44,314h 39m

D. Filtering ASS File to Generate Plain Text and Create Morphemes

As a post processing, we filter the ASS format files to generate plain language texts without meta-symbols. These plain texts are composed of Japanese sentences, and are finally divided into morphemes tagged by the part-of-speech

information. We use MeCab [10] as the morphological analyzer. Examples of a plain text and morphemes format file are shown in Fig. 2.

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
t*ochaku, 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

Fig. 2. An example of plain text extracted from an ASS file and its morpheme data processed using Mecab.

E. Converting TS Format to MP4 format

The size of a full-segment video file is very large. For example, the size of a TS file for 1 h of programming is about 2 GB. The total file size for the seven stations is over 200 GB daily. We therefore compress the file size by converting the original TS format file into an MP4 format file. The screen size of the MP4 format is also converted into 420 x 320 from the original 1920 x 1024. For this conversion, the freeware program, ffmpeg, is used. As a result of this process, the file size is compressed to one-tenth the original size on average.

F. Creating the Closed Caption TV Corpus

We built a large-scale closed caption TV corpus of over 70,000 TV programs collected from January 2013 to June 2014 [4]. Table I shows the numbers and hours of TV programs recorded every month. We have currently recorded a total of 70,037 programs, or 44,314 h and 39 min of programming. Each program is classified into at least one genre. The first genre listed is adopted as the primary genre to which the programs belong. Table II shows the total number of morphemes by genre.

The scale of our corpus at this point is 281,464,235 morphemes from 26,868,804 sentences. We can state that our corpus is presently the one of largest Japanese spoken corpus. In this research, we selected and used the sub-corpora from two genres: drama, labeled 'D,' and information and tabloid-style programs, labeled 'I.'

III. Method for Mining Recent Topics

Our purpose in this paper is to mine the recently popular topics from a large-scale TV closed-caption corpus. We focus on the words related to TV dramas in this paper. When a drama is especially successful, topics related to it also tend to become popular. We estimate that this phenomenon can be observed by comparing the word distribution between dramas and other genres.

TABLE II: The Total Number of Morphemes for Each Genre

Genre	Num. of Programs	Total Hrs (h:m)	Num. of Sents.	Num. of Morphes
A: Animation	5870	2202:16	1768621	13200874
S: Sport	2709	2296:35	1271335	15306196
C: Culture/ Documentary	8649	3833:22	1906741	22766430
D: Drama	7637	6761:37	4863814	41988885
N: News	13155	8551:16	3440839	61792042
V: Variety	11724	10203:41	7599892	63459148
F: Film	447	900:09	544867	4029948
M: Music	1222	818:33	314946	3406661
H: Hobby/ Educational	7281	2285:16	1741120	15698090
I: Information/ Tabloid Style	9993	5838:15	3187632	36925581
W: Welfare	986	421:45	225387	2818606
O: Other	30	22:24	3610	71774
Total	70037	44313:39	26868804	281464235

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 but also in other genres. On the other hand, words related to unpopular dramas might not appear in genres other than dramas because the drama went unnoticed by other TV programs.

We adopted the following process to discover topics we were interested in a while ago.

1) Counting the total number of words and the number of each word every month. We express the monthly distributions for each word by its monthly proportion.
2) Picking out words from the drama genre that appear in specific months. These words have the possibility of being the characteristic words in specific dramas.
3) Investigating how the words selected in Step (2) appear in other genres. It is expected that if one word is related to a popular drama, the distribution of the word in other genres is similar to the distribution in the drama genre. We find the words that have a similar distribution in other genres.
4) Checking whether the selected words from Step (3) can be considered to express the topics we were interested in a while ago. During this step, we must confirm the answer by checking texts that contain the target words in the corpus for the present.

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

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

where freq *i, j* refers to the frequency of the word *i* in month *j*.

The ratio *i, j* that is the proportion of word *i* in month *j*, is calculated with the Equation (2).

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

where *N* is the total number of words in the sub-corpus, and *Nj* is the total number of words in the *j-th* month in the sub-corpus. The *j* varies 1 to 18 because our corpus has terms of 18 months. Therefore each variable of frequency ratio for each word in a sub-corpus has 18 values.

In Step 2, we set three thresholds: the monthly frequency, the frequency ratio per month, and the word frequency. For the first threshold, we used a metric similar to the document frequency (DF) in the information retrieval domain. The DF for word *i* refers to the number of documents that contain the word *i*. It is considered that the DF represents the level of a word's specificity in the entire document set. We use a monthly frequency instead of the DF. If word *i* appears in all months of our corpus, the monthly frequency for word *i* is 18, because our corpus was collected over the 18 months. We seek words with monthly frequencies of less than nine months. For the second threshold, we seek biased words with a monthly frequency proportion of at least 10%. This 10% threshold means that the frequency of the word is biased approximately twice that of the average proportion, because the average frequency proportion for 18 months is approximately 5.5%. For the third threshold, we check the coverage of each word. In this research we set this threshold at 95% coverage. We sort all of the words in descending order and add up each frequency while the proportion of the added frequency for the entire size of the corpus is under than 95%. The word that is included in 95% of the coverage becomes an object for consideration. We mine for the words that satisfy all of these conditions.

In Step 3 we consider only the words selected in Step 2. For the other genres, we use the sub-corpus from information and tabloid-style programs, referred to as 'genre I.' We check whether the distribution for a selected word in Step 2 is similar to its distribution in genre I. We remove the words that have fewer similarities. We can set the similarity criteria. In this research, we use the following criteria to select word *i*.

1) The monthly frequency of word *i* from the I genre sub-corpus is the same or larger than it is in the D genre.
2) At least one of the monthly frequency proportions is over 10% in the genre I sub-corpus.
3) Pearson's correlation coefficient r between the D and I genres, equals 0.5 or higher. We calculate and check r for each word.

During Step 4, we check whether the words selected during Step 3 can be considered to express recently popular topics. We expect that the words related to a hit drama can be re-mined during these steps.

IV. Experimental Evaluation

To determine the effectiveness of our method for mining topics we were interested in a while ago, we performed two experiments. First, we examined how well our method performs with single word units. Second, we examined how well it performed with dual sequence word units, or bigrams.

A. Counting the Word Frequency

We counted the frequency of words in our corpus prior to the experiments. Table III provides the single-word lists by frequency according to genre, with 'D' representing dramas and 'I' representing informational programs.

The types of words in the D and I genres were 116,017 and 115,327, respectively. However, as shown in Table III, the total frequency for the top 10,000 words is over 95%. Approximately 10% of both genres' vocabularies have a coverage of 95% in the sub-corpuses. Therefore, we consider only the top 10,000 words by frequency in the D genre.

TABLE III: Single Word Lists by Frequency

	D: Drama (40813285)			I: Information/Tabloid Style (35914995)		
Rank	Word	Frequency		Word	Frequency	
1	◦	2654979	0.06	◦	2147758	0.05
2	*no*	1432079	0.10	*no*	1325869	0.09
3	*ta*	1086936	0.12	*te*	1104178	0.12
4	*te*	1072318	0.15	*ni*	927110	0.15
5	*ni*	968549	0.17	*ga*	863516	0.17
6	*wa*	934080	0.19	*desu*	859193	0.20
7	?	843797	0.22	*wa*	757987	0.22
8	*ga*	764225	0.23	*wo*	723276	0.24
9	*wo*	643202	0.25	*ta*	712036	0.26
10	*n*	593877	0.26	、	649488	0.28
10000	sum	38872417	0.95	sum	34385637	0.95

B. Experiment Involving Mining for Single Words

The first experiment we report assesses the topics that are mined with our method using the frequency of single words. First, we extracted candidate words from the results in Step (2), as described in Section III. In this experiment, we extracted 120 words as candidate words that are related to popular dramas. In the next step, we investigated how these 120 words appear in the I genre.

Next, we find words that have similar distributions in both genres, D and I. Fig. 3-Fig. 5 show the distributions of *Hanzawa*, eje and *Kanten*, respectively. Their values of Pearson's *r* are 0.932, 0832 and -0.208, respectively.

The former two words are judged to have similar distributions in both genres. *Hanzawa* is the hero of *Hanzawa Naoki* that was the great hit drama in 2013. In Japan, everybody knows him. *eje* is a part of a signature phrase *jejeje* that was frequently used in *Ama-chan* by the heroine of the drama. The third word is judged to do not have similar distribution. *Kanten* is a ager that is a general health food in Japan.

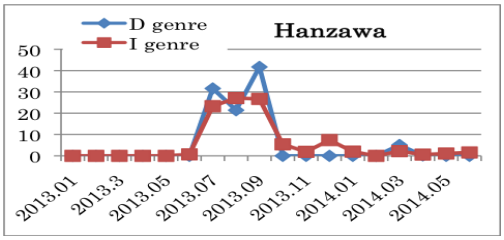

Fig. 3. A distribution of *Hanzawa* (*r*=0.932).

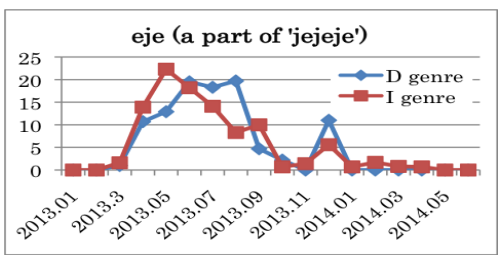

Fig. 4. A distribution of eje (*r*=0.832).

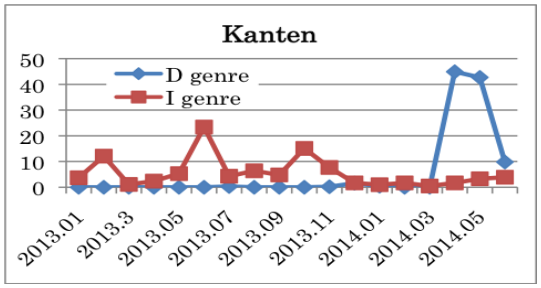

Fig. 5. A distribution of *Kanten* (*r*=-0.208).

Finally, we found 20 words that have similar distributions in both genres. Table IV lists the 20 words with their values of Pearson's *r*.

TABLE IV: THE EXTRACTED WORDS BY OUR METHOD

		Genre D			Genre I		
Word	*r*	Rank	TF	MF	Rank	TF	MF
Rei, 1, Ga	0.816	2485	1064	4	13103	101	5
Suzuka, 2, Am	0.844	2613	1017	8	11927	118	12
Ryo, 1, SR	0.501	3109	833	7	33661	16	7
Kikuta, 1, St	0.752	3570	712	5	29521	21	8
eje, 3, Am	0.832	3780	664	9	11412	126	14
Renko, 1, HA	0.922	4490	550	4	12583	108	5
GMT, 2, Am	0.794	4753	514	7	15633	75	8
*Hanzawa,*1, HN	0.932	4834	506	5	5461	400	12
Shiosai, 4, Am	0.734	4900	497	9	14726	83	11
Asaichi, 1, HA	0.707	4926	494	8	7150	265	18
Taisuke, 1, Go	0.568	5019	483	6	14311	87	6
Domyoji, 1, HD	0.527	5908	388	3	24180	32	6
kenji, 1, Ga	0.561	6527	340	4	39339	11	4
Itsuwa, 2, Rb	0.886	7361	289	3	26755	26	3
Himekawa, 1, St	0.813	7461	284	4	23241	35	7
Shozo, 1, Go	0.938	7932	259	6	16747	66	11
Morooka, 1, Go	0.958	8413	238	3	23335	35	3
*Yoshihira,*1, HA	0.732	8712	227	4	26931	26	4
Matsu, 1, Ar	0.599	9042	215	3	19053	52	3
Isejima, 2, HZ	0.999	9312	205	2	25818	28	2

In Table IV, 'TF' refers to the total frequency of a word, and 'MF' refers to its monthly frequency, as explained Section III. For example, the word *Hanzawa* appeared in 5 out of 18 months in the D genre and in 12 out of 18 months in the I genre. Numbers attached to words mean that 1 is names or nick-names of persons, 2 is organization names, 3 is signature phrases, and 4 is others. *Am*, *Go*, *HA*, *HN*, *Ga*, *St*, *Rb*, *HD*, *Ar* and *SR* refer 10 dramas *Amachan*, *Gochiso-san*, *Hanako-to-Ann*, *Hanzawa Naoki*, *Garasu no Ie*, *Strawberry Night*, *Roosevelt Game*, *HanayoriDango*, *Android*, and *Saiko-no-Rikon*, respectively.

The results of the first experiment show that four, three, three, and two words were related to the great hit dramas, *Ama-chan*, *Hanako-to-Ann*, *Gochiso-san*, and *Hanzawa Naoki*, respectively. The remaining eight words were related to six different dramas that were less popular.

From different viewpoint, fourteen words were names or nicknames of characters in dramas. Four words were parts of the names of organizations. Only one word was a part of a signature phrase from a heroine, and only one word was the name of the very popular hero in a drama.

It can be said that our method in the first experiment tend to yield parts of character names or signature phrases in the past dramas including greatly popular dramas. However the length of word seems too shorter as a topic for using directly. We should consider a method using phrases instead of words for topic detection.

C. Experiment Involving Mining for Bigrams

The second experiment we report assesses what topics are mined with our method using bigrams, or pairs of words. In the second experiment, we used bigrams instead of the single words in all processes, as stated in Section III.

In our corpus, the types of bigrams in the D and I genres were 2,473,561 and 2,318,971, respectively. The total numbers of bigram frequencies in the D and I genres were 45,677,768 and 39,102,669, respectively.

We extracted 788 bigrams as the candidate words that are related to popular dramas. In the next step, we investigated how the 788 bigrams that were extracted appear in the I genre. Fig. 6, 7 and 8 show the distributions of bigrams *Amano/Aki*, *j/ejeje* and *Rajio/Kyoku*, respectively.

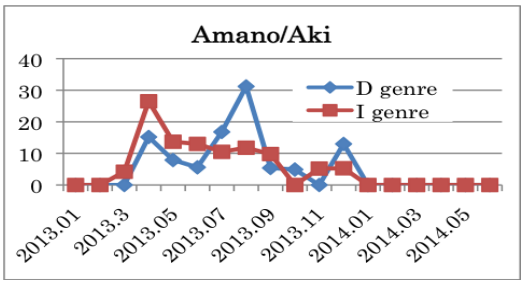

Fig. 6. A distribution of Amano/Aki (*r*=0.648).

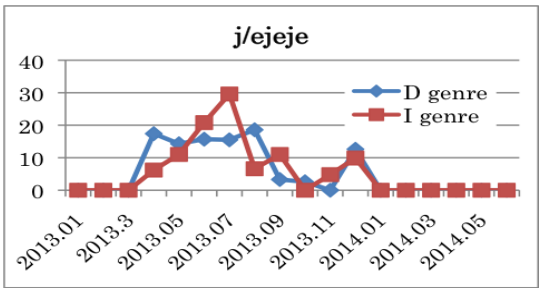

Fig. 7. A distribution of j/ejeje (*r*=0.726).

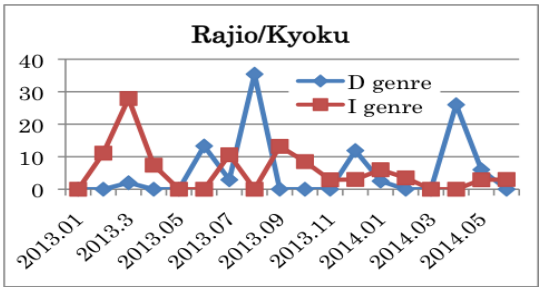

Fig. 8. A distribution of Rajio/Kyoku (*r*=-0.330).

The former two bigrams are judged to have similar distributions in both genres. *Amano/Aki* is the heroine of *Ama-chan* that was the great hit drama in 2013. In Japan, everybody knows her. Jejeje is a signature phrase of her and is also the winner of the keywords-of-the-year contest for 2013. The third bigram is judged to do not have similar distribution. *Rajio/Kyoku* is just a radio station. Their values of Pearson's *r* are 0.648, 0726 and -0.330, respectively.

Finally, we found the 119 bigrams that have similar distribution in both the D and I genres.

The extracted bigrams in the second experiment corresponding to 22 dramas that were different from the

results in the first experiment. 9 dramas out of 22 contained more than two bigrams. Table V shows titles of the 9 dramas and examples of bigrams.

TABLE V: THE EXTRACTED BIGRAMS AND DRAMAS

Title	N	Bigrams
***Ama-chan* あまちゃん	36	じ/ぇじぇじぇ(*gegege*), 天野/アキ(*Amano Aki*), 夏/ばっ(Natsu Ba),ばっ/ば(*Babba*),北/三陸(*Kita-Sanriku*),...
***Hanako-to-Ann* 花子とアン	20	花子/と(*Hanako to*),と/アン(*to* Ann),りょう/。,バーン/校長,...
***Gochiso-san* ごちそうさん	15	め/以子(*Meiko*), 悠/太郎(*Yutaro*),焼き/氷(*yaki-gori*),「/ごちそう(*gochiso*),...
**Gunshi-Kanbei* 軍師官兵衛	7	官兵衛/の(*Kanbei no*),黒田/官兵衛(*Kuroda Kanbei*),布/武,黒田/家,...
**Jun-to-Ai* 純と愛	7	愛/君(*Ai kun*), 純/は(*Jun wa*), 里/や(*Satoya*)
Android アンドロイド	6	沫/嶋(*Matsusima*),黎/士(*Reiji*),安堂/麻(*Ando Asa*),麻/陽(*Asahi*),嶋/黎,BOS/黎
***Hanzawa Naoki* 半沢直樹	6	BOS/半沢(*Hanzawa*), 庁/検査,大和田/常務(*Oowada Jomu*),伊勢島/ホテル(*Iseshima* Hotel)
Long good-bye	4	原田/保(*Harada Tamotsu*),原田/志津香(*Harada Shizuka*),増沢/磐
Roosevelt Game	3	沖/原(*Okihara*), 青島/製作所(*Aoshima/Seisakujo*) ,...

In Table V, 'N' refers the number of bigrams. Double asterisked titles and single asterisked titles refer great hit dramas and hit dramas, respectively. The results of the second experiment show that thirty-six, fifteen, twenty, and six bigrams were related to the great hit dramas, *Ama-chan*, *Hanako-to-Ann*, *Gochiso-san*, and *Hanzawa Naoki*, respectively. Two single asterisked hit dramas did not appear in the results of the first experiment. *Gunishi Kanbee* and *Jun-to-Ai* related to seven bigrams each. The remaining three dramas in Table V, *Android*, *Long good-bye*, and *Roosevelt Game* related to six, four and three bigrams, respectively. The remaining 15 bigrams were related to thirteen different dramas that were less popular.

From different viewpoint, 68 out of 119 bigrams were names or nicknames of characters in dramas. 24 bigrams were parts of the names of organizations. Only six bigrams were parts of signature phrases from main characters, and the remaining 21 bigrams were classified to others.

The result of the second experiment was similar to the result of the first experiment. Our method in the second experiment also tend to yield parts of character names or signature phrases in the past dramas including greatly popular dramas. The method using bigrams can extract more fragments of topics more than the method using single words. However the length of bigrams seems not enough and still shorter for a topic phrase. For future work, we should consider a method for making a chunk of words appropriate for the topic detection.

V. CONCLUSION AND FUTURE WORK

In this paper, we described methods for mining past topics we were interested in a while ago, from a closed caption TV corpus. The results showed that some words or bigrams related to the greatly hit dramas were extracted. It can be said that there is a possibility that a word having high correlation between the D and I genres is related to character names or signature phrases of dramas.

To improve the accuracy and availability of our method, we intend to pursue the following future projects to further our work in mining recent topics.

1) To compare the D genre with the other genres other than the I genre.
2) To count the total number of words or bigrams and the number of each word or bigram every week, rather than every month. We shall express weekly distributions of each word by the proportion of each week.
3) To use a longer unit instead of a single word or bigram.

REFERENCES

[1] L. Flowerdew, *Corpora and Language Education*, Palgrave Macmillan, 2011.

[2] J. Newman, H. Baayen, and S. Rice, "Corpus-based studies in language use, language learning, and language documentation," *Language and Computers Studies in Practical Linguistics*, Rodopi, 2011.

[3] K. Maekawa, H. Koiso, S. Furui and H. Isahara, "Spontaneous speech corpus of japanese," in *Proc. the Second International Conference on Language Resources and Evaluation LREC2000*, pp. 947-952, 2000.

[4] M. Mathioudakis and N. Koudas, "Twitter monitor: trend detection over the twitter stream" in *Proc. the ACM SIGMOD International Conference on Management of Data, SIGMOD 2010*, pp. 1155-1158, Indianapolis, Indiana, USA, June 6-10, 2010.

[5] N. S. Glance, M. Hurst, and T. Tomokiyo, "BlogPulse: Automated trend discovery for weblogs," in *Proc. WWW2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, 2004.

[6] J. Wang, X. W. Zhao, H. Wei, H. Yan, and X. Li, "Mining new Business opportunities: identifying trend related products by leveraging commercial intents from microblogs," in *Proc. the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1337-1347, Seattle, Washington, USA, October 2013.

[7] S. Nakajima, J. Zhang, Y. Inagaki, and R. Nakamoto, "Early detection of buzzwords based on large-scale time-series analysis of blog entries," in *Proc. the 23rd ACM Conference on Hypertext and Social Media (ACM Hyper- text 2012)*, pp. 275-284, June 2012.

[8] *Service Information for Digital Broadcasting System* (in Japanese), Association of Radio Industries and Businesses, 2009.

[9] T. Kudo, K. Yamamoto, and Y. Matsumoto, "Applying conditional random fields to Japanese morphological analysis," in *Proc. the Conference on Empirical Methods in Natural Language Processing EMNLP 2004*, pp. 230-237, 2004.

[10] H. Mochizuki and K. Shibano, "Building very large corpus containing useful rich materials for language learning from closed caption TV," in *Proc. the Association for the Advancement of Computing in Education (AACE) World Conference of E-Learning (E-Learn 2014)*, New Orleans, USA, November 2014.

Hajime Mochizuki was born in Japan. He obtained a PhD in information science from Japan Advanced Institute of Science and Technology in Ishikawa, Japan in 1999. He is currently an associate professor at the Institute of Global Studies, Tokyo University of Foreign Studies. His research interests include natural language processing and language learning system.

Dr. Mochizuki is a member of the Information Processing Society of Japan (IPSJ), Japanese Society for Information and Systems in Education (JSISE), and Association for the Advancement of Computing in Education (AACE).

Kohji Shibano is a professor at the Research Institute for Languages and Cultures of Asia and Africa, Tokyo University of Foreign Studies. His research interests include applied linguistics, language learning system, and information system. He serves as a convener of ISO/IEC 13249 SQL Multimedia and Application Packages committee (ISO/IEC JTC1 SC32/WG4).