Chinese Tag analysis for foreign movie contents

Lin, Xiao
Department of ISEE

Ito, Eisuke
Research Institute for Information Technology, Kyushu University : Associate Professor

Hirokawa, Sachio
Research Institute for Information Technology, Kyushu University

出版情報：Proceedings of IEEE/ACIS ICIS2014, pp.163-166, 2014-06-04. IEEE/ACIS
バージョン：
権利関係：

Chinese Tag analysis for foreign movie contents

Xiao Lin
Dept. of Information Science and Electrical Engineering,
Kyushu University
Fukuoka, Japan
lin.xiao.877@s.kyushu-u.ac.jp

Eisuke Ito and Sachio Hirokawa
Research Institute for Information Technology,
Kyushu University
Fukuoka, Japan
{ito.eisuke.523@m, hirokawa@cc}.kyushu-u.ac.jp

***Abstract*—Consumer Generated Media (CGM) is gaining huge popularity. The authors are particularly interested in the intercultural comprehension of movie contents made in foreign countries. This paper focuses on the website bilibili.tv as a test case to analyze how Japanese movie contents are watched in China. The authors analyze all tags and how foreign tags are introduced and translated into Chinese. They propose a simple statistical method to identify whether a word is a loanword or not, if the word is represented by Chinese characters. They also analyze the trends of tags in bilibili.**

Keywords—consumer generated media; CGM; movie sharing service; bilibili; tags; foreign word

I. INTRODUCTION

Consumer Generated Media (CGM) is gaining huge popularity. Particularly, the video-sharing websites such as YouTube (youtube.com) and Nicovideo (nicovideo.jp) are more popular among young people than television. We observe a similar situation in China, where youku.com, tudou.com, sohu.com and bilibili.tv are the most popular sites. The present paper concerns intercultural comprehension of movie contents made in foreign countries. As a test case, this paper analyzes how Japanese movie contents are watched in China. This paper focuses on bilibili.tv and the words used in those movies to represent foreign contents.

Bilibili is a video-sharing website based in China, where users can upload, view and add comments to videos. Mr. Xu Yi started a prototype website Mikufans.cn on June 26, 2009 [1,2], which was renamed as Bilibili and continues till the present day. Bilibili does not keep the movie contents on its own site, but instead provides meta-data to third party contents for the users. Most of the subtitles are in Chinese. They provide metadata not only for nicovideo, but also U.S. and Korean movies. The site is sometimes referred to as a parasite site.

They provide users' comments synchronize with movie reproduction. Users can keep favorite lists and can create a group of fans with the same interest. The most characteristic feature is a real-time commentary subtitle. These subtitles are called “danmaku” (literally bullets). The function originally came from nicovideo.com. Bilibili displays the movies from nicovideo and in many occasions displays the subtitles from nicovideo. Bilibili is influenced by nicovideo and vocaloid culture in which amateur users can create their own music by a singing voice synthesizer. In fact, the past website Mikufans used Hatsune Miku, a humanoid character voiced by a singing synthesizer which was popular in nicovideo and had a similar website structure to nicovideo.

The concept of “Cool Japan” was introduced around 2002 to gain broad exposure of Japanese pop culture, such as games, manga (cartoon), animation, J-POP and idols. Actually, it is reported that there are deep-rooted fans abroad with respect to games, manga, anime, J-POP and idols [3]. In order to know the use trend in China of Japanese animation contents, we started analyzing bilibili.

We have analyzed the nicovideo website before and know the current situation of nicovideo [4]. Bilibili has an affinity with nicovideo. Many contents of nicovideo are used in Bilibili. Bilibili is suitable for a trend survey of Japanese animation viewing in China. In the present paper, animation tags are analyzed as the beginning of Bilibili analysis. We focus on Japanese tags and how they are introduced and translated into Chinese. We pay attention to the new words and the coined words.

The rest of the paper is organized as follows. Section 2 analyzes the distribution of tag frequencies used in Bilibili. Section 3 concerns the foreign words and their Chinese translations. Section 4 concludes the paper and describes the further work.

II. FREQUENCY STATISTICS OF TAGS OF BILIBILI

We analyzed the tags of foreign animations. The analysis is based on the meta-data tagged on bilibili movies. We collected those meta-data submitted from November 2013 to January 2014. Each movie in bilibili is assigned with an identifier AV-ID that consists of characters (av) and numbers. We generated the numbers from 1 to 990,000 and succeeded to collect about 480,000 meta-data in HTML format. Table I shows detail of data.

TABLE I. COLLECTED METADATA FILES

Item	# of files	Total file size
HTML	909,398	19.24GiB
Valid HTML	480,257	14.03GiB

Fig.1 shows rough structure of a metadata and user’s comments for a movie. As we mentioned before, the movie

data are stored in another movie service site. Bilibili.tv only provides the metadata of the movie and posted user's data, such as tags, comments, number of viewers, given coins and points.

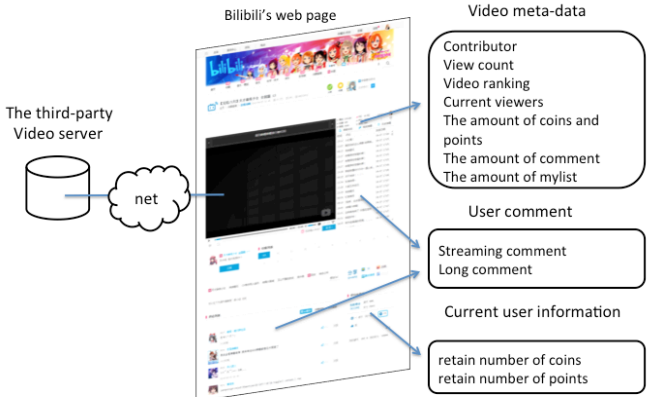

Fig. 1. Rough structure of a movie (medatada) page in bilibili.tv

Fig. 2. Number of replay, mylist, comments, etc. (av689970)

Fig.2 shows an example of meta-data, where we can see the title, the contributor, the view count, the amount of coins, the number of mylists (or bookmarked users) and the number of "danmaku".

The tags are written in the 2nd line from the bottom for each animation by the tele-viewer as shown in Fig.3. Viewers can assign at most 10 tags to an animation, which is the same as nicovideo. Fig.3 displays the meta-data for the movie whose id is "sm21443197" in nicovideo, which can be seen in the bottom line.

Fig. 3. Example of tags for a movie in bilibili.tv (av689970)

We extracted 345,140 unique tags from the collected meta-data. More than 250,000 tags were only assigned once to an animation. Table II shows the number of tags with respect to their frequencies.

TABLE II. NUMBER OF UNIQUE TAGS

Freq.	Num. of tags	Ratio (%)
1	249,825	72.4
2	34,013	9.9
3	14,595	4.2
4	8,579	2.5
More than 5	38,128	11.0
Total	345,140	100.0

III. TAGS FOR FOREIGN ANIMATION

This paper considers the influence of Japanese animation culture by analyzing Chinese tags assigned to the Japanese animations. Table III shows estimated user ratio of each countries in bibilibili provided by Alexa[5]. Most users are accessed from China, and they may be Chinese. But 40% users accessed from other nations.

TABLE III. RATIO OF USER COUNTORY

Country	Ratio
China	60.8%
Japan	11.1%
Taiwan	7.1%
United States	6.3%
Hong Kong	4.9%
Macao	3.4%
France	1.7%
Belgium	1.6%
Australia	0.6%
Switzerland	0.5%

A. *Notation of Foreign Language in Chinese*

By rough survey of the bilibili movies, it turned out that most Japanese movies are assigned the words of Japanese origin. Some of them are used as original Japanese words or with a little bit of modification. Some words are translated into Chinese words. The literal translation from Japanese "の" to Chinese "的" is an example of such a modification. "进击的巨人" is a modification of "進撃の巨人".

There are three kinds of modifications:

(a) based on the word sound,

(b) based on the word meaning and

(c) based on the sound and meaning.

The literal translation is widely used not only for animations but for the names of companies, services and products as well.

B. *Classification of Bilibili Tags based on Character Code*

Morphological analysis would be a method for categorizes the words. However, the tags, which we analyze in this paper, are one word or a short phrase. Morphological analysis is not appropriate for them. To categorize tags, we used a simple method using characteristics of UTF8 character encoding. All tags in bilibibli are written in UTF8 character encoding. It is possible to which language letters by specification of UTF8 encoding. Fig.4 shows UTF8 code characteristics for ASCII code and Japanese *hiragana* and *katakana* code. We found that

the first bit of ASCII character is 0, and the first byte of Japanese *hiragana* and *katakana* is always be $(E3)_{16}$. Then, it is easy to classify ASCII character only tags, and at least Japanese *hiragana* or *katakana* letter including tags.

Letter	Unicode (Hex)	UTF-8 (Binary)
A	65	01100101
林	67 97	11100110 10011110 10010111
あ	30 42	11100011 10000001 10000010

We can find the first byte of hiragana and katakana is always be $E3_{16}$ $(11100011)_2$

Fig. 4. UTF-8 code characteristics for Japanese *hiragana* and *katakana*.

Let T be the set of all tags contained in meta-data of bilibili animation. Those tags are written in UTF8 character codes, which distinguish the character of each country. We used this feature of UTF8 code to separate T into four groups T_1, T_2, T_3 and T_4. Table 3 displays the feature of each group and the number of tags. Note that T_i and T_j don't have common tags.

TABLE IV. NUMBER OF TAGS IN EACH TAG SET

Symbol	Description	# of tags	Ratio
T	All tags	345,140	100.0%
T_1	Include at least one Japanese "*Hiragana*" or "*Katakana*"	27,371	7.9%
T_2	English Alphabet only	43,913	12.5%
T_3	Chinese *Hanzi* only	268,575	77.8%
T_4	Num. and Symbol only	5,281	1.5%

C. Foreign Words in Chinese

We want to separate the tags into the original Chinese words and foreign words. The words in the group T_1 and T_2 are foreign words. It is not trivial to tell if a word in the group T_3 is an original Chinese word or a foreign word in Chinese, since all characters are Chinese character. We selected the typical Chinese characters used in representing a foreign word. Fig.5 shows 100 characters we chose together with their pronunciation (*PinYin*). We denote the set of these 100 characters by K.

阿:a, 拉:la, 亚:ya, 维:wei, 萨:sa, 斯:su, 罗:luo/lo, 巴:ba, 卡:ka, 兰:lan, 肯:ken, 特:te, 利:li, 姆:mu, 哈:ha, 克:ke, 纳:na, 西:shi, 尔:er, 基:ji, 乌:wu, 塞:sai, 安:an, 俄:e, 加:jia, 达:da, 尼:ni, 圣:sheng, 奥:ao, 苏:su, 弗:fu/fo, 威:wei,伦:lun, 蒙:meng, 诺:no, 坦:tan, 塔:ta, 兹:zi, 摩:mo, 丹:dan, 黎:li, 索:so, 雅:ya, 艾:ai, 凯:kai, 班:ban, 托:tuo/to, 圭:gui, 布:bu, 埃:ai, 伊:i, 格:ge, 勒:le/lei, 法:fa, 库:ku, 洛:luo/lo, 本:ben, 马:ma, 哥:go/ge, 昂:ang, 赞:zan, 桑:san, 茨:ci, 莱:lai, 敦:dun, 瓦:wa, 士:shi, 伯:bo, 米:mi, 麦:mai, 卢:lu, 雷:lei, 夫:fu, 曼:man, 纽:niu, 昆:kun, 里:li, 朗:lang, 赫:he, 波:bo, 拿:na, 厄:e, 廷:ting, 喀:ka, 康:kang, 菲:fei, 舍:she, 泽:ze, 宾:bin, 瑟:se, 瑞:rui, 莫:mo, 涅:nie, 犹:you,温:wen, 丘:qiu, 德:de, 森:sen, 顿:dun, 霍:huo/ho

Fig. 5. The set K. Frequently used Chinese *Hanzi* letters for foreign words

We used the ratio of the characters of K in a word in T_3 to approximate the level of a foreign word. If a word contains no characters of K, then the word is likely a word of Chinese origin. If a word contains many characters of K, then the word may be a foreign word. Depending on the number i of the characters of K in a word, we classify T_3 into S_i's, where S_i is the set of words in T_3, which contain more than i occurrences of characters of K.

We made a program to extract the words of S_i from T_3. Then we obtained Table V which shows the number of tags in S_1, S_2, ..., S_5. We manually checked if each of word in S_i is a foreign word or not. Table III shows the result which confirms the hypothesis that the more often K words are used in a tag, the more likely the word is a foreign word.

$$S_i = \{\, t \mid t \text{ in } T_3,\ t \text{ includes at least } i \text{ character(s) in } K \,\}.$$

TABLE V. NUMBER OF TAGS IN EACH TAG SET

Symbol	# of tags	Ratio for T_3
S_1	32,206	12.0%
S_2	6,519	2.4%
S_3	2,895	1.1%
S_4	992	0.4%
S_5	429	0.2%

IV. ANALYSIS OF FREQUENT TAGS

This section analyzes the frequent tags of T_1, T_2 and T_3 in detail.

A. Tags of Japanese Origin

Fig.4 shows the top 30 frequent tags in T_1. The words of T_1 contain at least one *hiragana* or *katakana* Japanese character. So, we expected that the tags in T_1 would be Japanese words or phrases. Our expectation was correct as Fig.4 shows. There are many original Japanese words used in the Japanese website nicovideo. Most of the tags in T_1 are the names of performers and characters of a program or the titles of music.

The tags, which are used to represent the opinions such as "should be evaluated better" and "who benefits" are rarely used. Those tags are written in Chinese words.

初音ミク, 歌ってみた,镜音リン, 巡音ルカ,镜音レン, 重音テト, 真夏の夜の淫梦, 迷の感动, 踊ってみた, 日刊妹俺の嫁, 神威がくぽ, 迷の高产, 红莲の弓矢, 波音リツ, 结月ゆかり, レトルト, 赤ティン, 舰これ, 白金ディスコ, 猫村いろは, 独りんぼエンヴィー, まふまふ, アイドルマスター, 合唱シリーズ, 镜音リン・レン, サリシノハラ, そらる, VOCALOID→UTAU カバー曲, 演奏してみた, ミクオリジナル曲

Fig. 6. Top 30 tags in T_1 (Including Japanese Characters).

B. Alphabetic Character Tags

Fig.5 displays the top 30 tags in T_2. The words in T_2 consist only of alphabetic characters. Most of them are, as we expected, an English word, an acronym or a coined word originated in US. Other samples of T_2 tags are words of Japanese origin such as VOCALOID, Arashi, cosplay and AKB48.

VOCALOID, LOL, MMD, MINECRAFT, MUGEN, DOTA, APH, DOTA2, UTAU, GUMI, MIKU, DNF, KAITO, OSU, BGM, OP, CS, WOT, PS3, IA, ARASHI, GALGAME, YOUTUBE, WOW, MC, MAD, PV, COSPLAY, AKB48, PSP

Fig. 7. Top 30 tags in T_2 (Alphabet) .

C. Chinese Character Tags

T_3 is the main target of our analysis. We expected to understand how foreign cultures are adapted in the Chinese subtitles of animations. The distinction of words of foreign origin is relatively easy in Japanese, since they are mostly written in Japanese *katakana* characters. In Chinese, every word is written in Chinese *Hanzi* characters even if it is a word of foreign origin. Fig.6 shows the top 30 tags in T_3.

东方，英雄联盟，实况，翻唱，解说，东方MMD，游戏，原创，洛天依，坦克世界，鬼畜，搞笑，进击的巨人，三国杀，元首，初音，我的世界，娱乐，福利，日剧，优酷，音乐，黑塔利亚，星际2，银魂MMD，游戏实况，高清，吐槽，游戏王，卖萌

Fig. 8. Top 30 tags in T_3 (Chinese *Hanzi* only) .

There are words of Japanese origin words such as “东方” and “初音”, which do not have direct translation. There are many entertainer names below the top 30. On the other hand, we found some Chinese original words that have not translation in Japanese. There are many general words such as “entertainment” in T_3.

D. Chinese Characters Representing Foreign Words

Fig.3 shows the set K of 100 Chinese *Hanzi* characters, which are used often to represent foreign words. We selected those Chinese characters based on the situation where they are used. They are used to denote the names of European or American people and English words.

Table V explains the effect of K to guess the foreign origin words. Chinese words with a few K characters are Chinese original words. A word with more than three occurrences of K characters can be identified as a foreign word. However, the number of such words is not large.

A dictionary and corpus would be useful to identify the foreign words. In Table V and the set K, we consider one character to identify the foreign words. Statistic analysis of the sequence of characters in foreign words would be applicable.

A metrical technique method would be applicable in selecting the set K of Chinese characters. In Wikipedia, English words and Japanese words are explained in Chinese. The frequency analysis of the occurrences of Chinese characters would be a reasonable approach to consider the candidates for K.

V. Conclusion

The present paper is the first step to analyze the intercultural effects of web media. We collected meta-data of the bilibili website. We analyzed the tags assigned to foreign movies to guess how foreign words are represented in Chinese characters. 100 Chinese characters were chosen to identify the words of foreign origin.

Further work is necessary. Discovering the translation pair of words will be possible from the set of tags for the same movie, if the tags contain both Japanese and Chinese words. If a Chinese word co-occurs frequently with a Japanese name of hero or heroin, then the Chinese word would be good candidate of his/her name in Chinese. If a video of nicovideo can be seen in bilibili as well as in nicovideo, the Japanese tabs and the Chinese tags could be used as a source to constructing a dictionary.

Time series analysis of the frequency of tags will also be interesting. It will be possible to see how movies of other cultures are accepted among people, if we focus on tags assigned to foreign movies. The trend analysis of tags for movies will be applicable as log as we can obtain the meta-data. By considering the correspondence of tags of the same movie in different languages, the spread of culture and the comparison of the attitude of people will be possible.

Acknowledgment

This work was supported by KAKENHI 23500299.

References

[1] Bilibili (Jan.20,2014). In Wikipedia: The Free Encyclopedia. Retrieved from http://zh.wikipedia.org/wiki/Bilibili

[2] Biliili (Jan.20,2014). In Wikipedia: The Free Encyclopedia. Retrieved from http://ja.wikipedia.org/wiki/%E5%97%B6%E5%93%A9%E5%97%B6%E5%93%A9

[3] T. Sakurai: “The Chinese women who very like Japanese products”, PHP, ISBN-10: 4569812422, 2013. (in Japanese)

[4] N. Murakami, E. Ito, “Emotional video ranking based on user comments,” Proc. of iiWAS2011, ACM, pp.499-502, 2011.

[5] Alexa (April.10,2014). http://Alexa.com/siteinfo/bilibili.tv