動画サービスにおける外国動画に付与されたタグの分析

林 驍　　伊東 栄典
九州大学システム情報科学研究府，九州大学情報基盤研究開発センター
lin.xiao.346@s.kyushu-u.ac.jp　ito.eisuke.523@m.kyushu-u.ac.jp

1. はじめに

近年，YouTube (youtube.com)，ニコニコ動画 (nicovideo.jp) などの利用者投稿型動画共有サービスが人気である。中国でも優酷(优酷 youku.com)や土豆 (tudou.com)，搜狐（sohu.com）および嗶哩嗶哩動画 (哔哩哔哩动画 bilibili.tv)が人気である。我々は，日本の動画コンテンツの中国での利用度合いに興味を持っている。本論文では，嗶哩嗶哩動画(以下 bilibili と略記) を対象に，中国からみた外国コンテンツに付随する語を分析する。

Bilibili は，AcFun の会員である⑨bishi 氏が 2009 年 6 月 26 日に設立したファンサイト「Mikufans」を前身としている[1,2]。Mikufans は 2010 年 1 月 24 日から現在のサイト名「嗶哩嗶哩 (bilibili) 」に改名され，現在まで続いている。Bilibili は動画の実データを自サイトに保存しておらず，別の動画サイトが保持する動画を使い，それに利用者向けのメタデータ（主に中国語）提供している。また他サイトにある動画再生に同期して，弾幕と称する視聴者コメント（評論）を動画面上に表示する機能や，動画についての利用者のお気に入り情報の保持，およびファン向けのグループ作成機能などを提供している。これらの機能から，Wikipedia[1,2]には Bilibili は寄生型サイトと表記されている。

Bilibili は日本のニコニコ動画およびニコニコ動画から発した Vocaloid 文化に強く影響されている。前身サイトの Mikufans の Miku は初音ミクであるし，またサイトの構成がニコニコ動画に類似している。動画再生に同期しての動画面上への弾幕コメント（評論）表示は，ニコニコ動画のコメント表示と同じ機能である。更に，ニコニコ動画サイトへの投稿動画を bilibili 内で表示する場合も多い。ニコニコ動画以外にも，米国や韓国系の動画も多い。

近年，日本の近代文化であるゲーム・漫画・アニメ， J-POP・アイドルなどのポップカルチャーをクールジャパン推進の一部として海外に積極的に展開する動きがある。既にゲーム・漫画・アニメ・J-POP・アイドル等は海外に根強いファンが居る状況であるとの報告もある[3]。実際，筆者を含む中国の若者には，日本アニメを見て成長した者が多い。

日本動画コンテンツの，中国での利用動向を知るために，bilibili 動画のタグを解析した。我々は過去にニコニコ動画サイトの解析をしており[4]，ニコニコ動画に詳しい。Bilibili はニコニコ動画と親和性があり，ニコニコ動画のコンテンツを多く用いるため，日本動画の動向調査に適している。Bilibili 解析の手始めとして，動画タグを解析する。特に，中国からみて外国語となる日本語のタグ，特に新語や造語が中国語にどのように翻訳されているかに注目する。これらの知ることで，アニメなどのニッチな文章の機械翻訳などに貢献できると思われる。

本論文の構成を述べる。2 節では bilibili 動画のタグについて，出現頻度等の基礎分析結果を述べる。3 節では中国での外国語の扱いについて考察する。最後に 4 節でまとめと今後の課題を述べる。

2. Bilibili 動画タグの基礎分析

外国動画に付けられたタグを分析するために，先ず bilibili 動画に投稿されている動画メタデータを 2013 年 11 月から 2014 年 1 月の間に収集した。Bilibili では動画に AV ID と呼ぶ識別子（av+数字）を付与する。AV ID は文字 av と数字で構成されるため，数字部分を 1 から 99 万まで 1 つずつ増やしてアクセスした所，約 48 万個の動画についてメタデータ（HTML 形式）を収集できた。

図 1 に示すように 1 つの動画のメタデータには，題名，投稿者（登録者名）の他に，動画再生回数（播放），コイン数（硬币数量），マイリストあるいはブックマーク数（收藏），弾幕コメント数（弾幕）の情報が有る。

図 1　Bilibli での動画利用状況情報 (av689970)

図 2 に示すように，動画にはタグが視聴者により付与されている。ニコニコ動画と同様に，bilibili でも最高 10 個のタグを動画に付与できる。図 2 の動画は，ニコニコ動画で「sm21443197」の識別子を持つ動画であるため，その情報がタグに記載されている。

図 2　Bilibli でのタグ情報 (av689970)

集めた全動画メタデータからタグを抽出した。一意なタグは 345,140 個あった。出現頻度で分けた所，1 回しか出現しない（1 つの動画にしか付与されていない）タグが 25 万個あった。

表 1 一意なタグの個数

頻度	タグ数	割合
1 回	249,825	72.4%
2 回	34,013	9.9%
3 回	14,595	4.2%
4 回	8,579	2.5%
5 回以上	38,128	11.0%
総数	345,140	100.0%

3. 外国動画に付与されたタグの分析

本研究では，中国からの外国動画についての分析を行うことで，日本の動画文化が与える影響を知りたいと考えている。

3.1. 中国における外国語の表記

Bilibili の動画をおおまかに調査したところ，日本由来の動画には，日本由来の単語がタグ付けされる場合が多い。それらは，日本語そのまま，あるいは少しだけ変更されて使われている場合と，中文漢字の単語に変換される場合がある。中文漢字への置換えは，(a)音を当てたもの，(b)意味での置換え，(c)音と意味を考慮した置換えがある。外国語の，中文漢字への置き換えは，動画にかぎらず，企業名やサービス名，製品名など，多くの分野で行われている。

例えば，(a)の例では，「ガンダム」を「高达」，「フランドール・スカーレット (Frandle Scarlet)」を「芙兰朵露·斯卡雷特」などがある。 (b)の例では「イカ娘」を「乌贼(烏賊)娘」，「らき☆すた」を「幸運星」などがある。(c)の例では，「まどか」を「馒头卡」と表記し，音を似せて，かつ意味を含めた当て字を作成している。「進撃の巨人」を「进击的巨人」として「的」だけ入れるものも多い。なお，「まどか」は，「馒头卡」以外にも「小圆脸」「圆神」などの単語になっており，これらを同義語として識別する必要がある。

3.2. Bilibili タグの分類

Bilibili の動画メタデータに含まれる一意なタグの集合を T とする。全タグの文字コードは UTF8 であるため，各国文字のコード特性を利用して，T を 4 つ(T_1～T_4)に分けた。各 T_i の意味およびタグ数を表 2 に示す。なお，$T_i \cap T_j = \phi$ $(i \neq j)$である。

表 2　文字コードによるタグ分類

表記	説明	タグ数	割合
T	全タグ	345,140	100.0%
T_1	平仮名片仮名を 1 文字以上含む	27,371	7.9%
T_2	全て英字(alphabet)	43,913	12.7%
T_3	全て中文漢字	268,575	77.8%
T_4	記号・数字	5,281	1.5%

3.3. 中国語での外来語

次に全て中文漢字で構成されたタグについて，外来語由来の単語と，中国由来の単語に分けることにした。単語の分類には形態素解析を使う方法も考えられる。しかし本研究が対象とするものは，タグとして用いられている単語または短いフレーズであるため，形態素解析は適さないと考え，単純な方法を適用することとした。

中国で，外来語の表現に用いられやすい漢字を，著者が人手で抽出した。思いついた外来語への当て字用漢字を 100 個選んだ。この中文漢字集合を K とする。表 3 に我々が選んだ K の漢字と，その発音 (PinYin) を示す。ただし四声は省いている。

この外来語用当て字集合 K を用いて，表 2 の T_3 集合のタグ（全て中文漢字から成るタグ）から，外来語由来のタグを抜き出す処理を行った。T_3 か

ら抜き出された後のタグ集合を S_i とする。S_i の定義を以下に示す。

$S_i = \{ t \mid t \in T_3, t$ は K の漢字を i 個以上含む $\}$.

S_i の定義に則したタグの抽出プログラムを作成し，T_3 からタグを抜き出した。抜き出したタグ集合 $S_1 \ldots S_5$ のタグ数を表 4 示す。

表 3　外国語に当てる中文漢字集合 K（100 個）

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

表 4　K の漢字を含むタグ数

集合	タグ数	T3 との割合
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%

次に，S_i の各タグについて，本当に外国語由来の語であるか，そうでないかを人手で調べた。調べた精度(precision)を表 4 に示す。

表 5　S_i の外国語由来の単語率

タグ集合	精度
S1	15.7%
S2	79.7%
S3	97.7%
S4	99.7%
S5	100.0%

4. 考察

4.1. 日本由来のタグ

表 4 の T_1, T_2, T_3 について，タグの傾向を分析する。最初に T_1 を分析する。

T_1 で，出現頻度上位 30 位のタグを表 6 に示す。T_1 は平仮名か片仮名を 1 文字以上含むタグであるため，日本語の単語やフレーズを使うタグと想定していた。表 6 に示すように，出現頻度の高いタグは，日本のニコニコ動画サイトで，動画に付けられたタグそのもののタグが多い。T_1 のタグは動画の内容を的確に表す，出演者（キャラクター）や音楽題名が多い。しかし，ニコニコ動画で良く見られる「もっと評価されるべき」「誰得」などの，動画を直接説明しない単語は少ない。動画を直接説明しないタグは，中国語で表現されている。

表 6　T_1 の出現頻度上位 30 個

初音ミク，歌ってみた,鏡音リン，巡音ルカ,鏡音レン，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，

4.2. 英字タグ

表 7 に，T_2 のうち出現頻度上位 30 位のタグを示す。英文字だけから成るタグは，想定していたとおり，英単語や米国由来の略語・造語が多くふくまれている。他にも VOCALOID や Arashi(嵐)，cosplay（コスプレ），AKB48 などの日本由来の単語も多いことが分かった。

表 7　T_2 の出現頻度上位 30 個

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

4.3. 中文漢字タグ

最後に T_3 の分析を述べる。T_3 は本研究の主たる対象で，中国における外国語動画の扱いを見るものであった。日本では外国語をカタカナで表記するため，外来語か否かの判別は比較的容易である。中国語の場合，漢字のみで表現するため，外来語の表現が独特である。

表 8 に T_3 のうち出現頻度上位 30 位のタグを示す。

表 8　T_3 の出現頻度上位 30 個

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

表 8 を見ると，東方や初音などの日本由来の漢字単語が含まれていることが分かる。表 8 には無いものの，日本の芸能人の人名（漢字のみのもの）も多く含まれている。

一方，卖萌（売萌，「あざとい」に近い意味）のような中国由来の言葉もある。T_3 には娯乐（娯楽）のような一般単語も多い。

4.4. 外国語の表現に使う漢字

表 3 には，我々が選んだ，中国で外国語表現に使いやすい漢字集合 K を示している。これらは欧米の人名や，英単語を表現する際に用いられやすいものとして選んだ。

この K の効果について考察する。表 5 を見れば分かるように，K の漢字を複数回使うか否かという単純な方法では，外国語判別の効果は低い。K の漢字を 3 個以上含む場合，ほぼ外国語であると判断できるが，K の漢字を 3 個以上含むタグは少ない。

タグが外国語であるか，否かを判定するには，辞書やコーパスが必要かもしれない。統計的解析で外国語の場合に連続しやすい漢字を見つけ，それを判定に使うのも良いかもしれない。

また，今回 K の漢字選出には計量的な手法を用いていない。Wikipedia には欧米の語や日本の言葉に対して，対応する中文（中国語）での説明も有る。これらのデータを用いて，頻出漢字解析を行うことで，より有用な，中国で外国語に当てる際に用いる中文漢字集合を選出可能であろう。

5. おわりに

日本由来または米国由来の動画の，中国での閲覧状況調査を大目標として，中国での外国語動画に付与されるタグの状況を調査している。中国における日本を含む海外動画の影響を調査するため，bilibili 動画を対象にタグの解析を行った。収集した動画メタデータから抜き出したタグの傾向を分析した。

本稿は第一段階の分析結果報告であり，今後多くの詳細分析が必要である。１つの動画タグ集合に，日本語と中国語が混ざる場合，翻訳候補を統計的に算出可能であろう。例えば「まどか」と「馒头卡」の共起頻度がたかければ，音の近さを計算することで，単語の翻訳候補を算出可能であろう。他にも，ニコニコ動画の 1 つの動画が bilibili でも閲覧可能な場合，ニコニコ動画側で付与されたタグ（日本語）と，bilibili 動画で付与されたタグ（中国語）の関係を解析すれば，半自動的に翻訳候補を提示可能である。

最後に，本稿で述べた外国動画のタグ傾向分析から，日本の動画が中国で閲覧されている状況や，米国動画の中国での動向，逆に中国動画の日本や米国での状況など，多くの国から発信されたコンテンツが世界に広まる様子を知ることが可能だと思われる。最終的には，動画の利用度合いから，各国の文化の影響度合いを解析していきたい。

文　献

[1] Bilibili (Dec.20, 2015). In Wikipedia: The Free Encyclopedia. Retrieved from http://zh.wikipedia.org/wiki/Bilibili

[2] 嗶哩嗶哩 (Dec.20, 2015). 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] 櫻井孝昌: 日本が好きすぎる中国人女子，PHP 新書, ISBN-10: 4569812422, 2013.

[4] Naomichi Murakami, Eisuke Ito, Emotional video ranking based on user comments, Proc. of iiWAS2011, ACM, pp.499-502, 2011.