FKC Corpus: a Japanese Corpus from New Opinion Survey Service

Kensuke Mitsuzawa†, Maito Tauchi†,
Mathieu Domoulin†, Masanori Nakashima†, Tomoya Mizumoto‡
†Fuman Kaitori Center
†6-5-1 Nishi-Shinjuku, Shinjuku-ku, Tokyo 163-1333, Japan
‡Tohoku University
‡6-6-05 Aramaki Aza Aoba, Aoba-ku, Sendai, Miyagi 980-8579, Japan
† {kensuke_mitsuzawa, maito_tauchi, domoulin_mathieu, masanori_nakashima}@fumankaitori.com
‡ tomoya-m@ecei.tohoku.ac.jp

Abstract

In this paper, we present the FKC corpus which is from Fuman Kaitori Center (FKC). The FKC is a Japanese consumer opinion data collection and analysis service. The main advantage of the FKC is the system that awards greater points to user input containing more information, which encourages users to input categorical information. Thanks to this system, the FKC corpus has consumers' opinions with abundant category and user demographics, and is considered to serve multiple NLP tasks: opinion mining, document classification, author inferring and sentiment classification. The FKC corpus consists of 254,683 posts coming from 25,092 users. All posts are checked by annotators who are working for the FKC in crowdsourcing. The posts in the FKC corpus mainly comes from mobile devices, and one third of them are about products or events related to daily life. We also show some correlations between point incentive and users' motivation which keeps posting their opinions with abundant category information.

The FKC corpus is available under an original license of the FKC. Currently, the FKC gives permission to use directly, thus, those who hopes to use the FKC corpus needs to send request to first author.

Keywords: Social Media, Corpus construction, Crowdsourcing

1. Introduction

Public datasets extracted from the web are a popular data resource for NLP research. This is especially true for modern NLP research which makes increasing use of machine learning for such research applications as document classification (Boley et al., 1999; Schenker, 2003), sentiment classification (Zhang et al., 2015), opinion mining (Orimaye et al., 2012), and author inferring (Mukherjee and Liu, 2010; Nguyen et al., 2011).

There are several issues that researchers commonly face when using many of the public datasets made of information on the web. First, these datasets are often noisy. They require time-consuming pre-processing before they can be used. Second, these data resources tend to be lack in contextual information (i.e. metadata) such as author profile, likewise class metadata can be inconsistent. Thus, analysts or researchers must often manually label their data before use, as in (Noll and Meinel, 2008).

In this paper, we introduce a novel Japanese language corpus. This is extracted from data accumulated by *Fuman Kaitori Center* (FKC)[1],which is a Japanese consumer opinion data collection and analysis service opened in 2015. "*Fuman*" means dissatisfaction in Japanese. The core concept of the FKC is to collect consumers' negative opinions about companies and their products or their services in exchange for a small monetary reward. This monetary reward is exchangeable with a gift card which is able to be used in an electronic commerce service. As a running web service, the FKC is accumulating data at the rate of 5-10,000 posts a day as of mid 2015. On the business side, the FKC offers an analytics dashboard and custom reports to whom wishes to know opinions on specific products or services as shown in Figure 1.

Figure 1: Analytics service from the FKC. Business users are able to check latest statistics with an Analytics dashboard (Left), to check suggestions from data with an Analysis Report (Right)

Considering the FKC corpus as dataset for NLP tasks, the FKC corpus has several major advantages. First, the corpus includes metadata such as user profile information accompanying the posts' textual content. In addition, the corpus is less noisy than other comparable public datasets, and the corpus is more focused, only including relatively short, negative opinions. Secondly, the FKC corpus is collected from a live service and is thus growing every month, making possible research applications that require time-series data. We believe that the FKC corpus can be a useful data source for a great variety of NLP tasks.

In this paper, we first show related dataset and platforms in Section 2. Next, a brief introduction of the FKC is presented in Section 3. The Section 4 describes statistics in the

[1]http://www.fumankaitori.com

FKC corpus. The Section 5 shows correlations between a point incentive system of the FKC and users' motivation. We give some examples of NLP application in Section 6. Finally, we make the conclusion in Section 7.

2. Similar datasets for NLP tasks

2.1. Twitter

Twitter is a global-scale SNS service used by people for sharing short thoughts, opinions, and observations in near-real-time either publicly or to a private group of "followers". For several years now, Twitter has been a popular resource for NLP-related research (Sasa et al., 2010; Pak and Paroubek, 2010). But using text data extracted from Twitter causes some problems. For example, it is hard to classify tweets by their topics, moreover user demographic information tends to be unknown. To be fair, there is some metadata in Twitter, like user's age and location for user demographic information, and hashtags and geo-tagging for tweets. But, as we mentioned, the user demographic information tend to be unknown because there are less merits to fill in for ordinary users. Filtering on hashtags might miss relevant posts without hashtag, otherwise that might include unrelated posts that have the hashtag spuriously, making the dataset potentially very noisy. Therefore, it is laborious task to make clean data from Twitter.

While the FKC corpus is significantly smaller than data extracted from Twitter, it is more focused, with more well-defined categories and topics. Moreover, the FKC corpus has user profile information which adds demographics to the analysis.

2.2. Youtube

Youtube allows its users to post comments for each video. These comments can be used as a relevant data source for such tasks as opinion mining. For example, Uryupina et al. (2014) uses the posted comments from promotional videos as a dataset for opinion mining. While their dataset includes some metadata, such as the video URLs and external links to related products, it does not include any user profile information. In addition, the comments are not categorized, therefore it includes some irrelevant comments, making the dataset rather noisy.

Compared this dataset with the FKC corpus, it has the advantages of user profile and less noiseness.

2.3. Rakuten Data

Rakuten, which is one of the largest e-commerce company in Japan, makes several dataset available[2]. The one of their dataset, the Rakuten Ichiba dataset includes product data for over 150 million items as well as over 64 million user reviews about these items. Moreover, it is notable in a lot of metadata, such as user profile and review rating.

While the Rakuten dataset has review text and a lot of metadata, their reviews are limited to a specific domain. For example, reviews in Rakuten Ichiba data are only for products, also for shop owners who sell products inside Rakuten Ichiba. On the other hand, the FKC collects opinions without domain limitation such as "human relationship", "public service" and "politics", which are useful for analysts or researchers who carry out public opinion analysis.

2.4. MPQA opinion corpus

MPQA opinion corpus is annotated dataset which is consisted of 506 documents mainly from news articles. This dataset is open at website[3] and dataset description is in Wiebe and Theresa Wilson (2005).

MPQA dataset is worthy because of its wide variety of metadata information. In this dataset, *private state* (Ex. emotion, sentiment, belief, speculations etc.) metadata is annotated for words and phrases. Moreover, metadata is categorized by its expression level which is from direct expression to indirect expression. And the text is well-formatted style because its documents are mainly from news articles.

Although MPQA is good for its rich annotations, the dataset contains static information, from which public opinion is difficult to determine. The FKC corpus comes from lively posts, thus we are able to know public opinions from it.

3. A brief introduction of FKC

Fuman is a Japanese word which is usually translated into English as discontent or dissatisfaction. It can be tied to various negative feelings such as anger, sadness, disappointment, frustration and so on. Most kinds of fuman are posted to the FKC by consumers when they are faced with a recent unsatisfactory experience from a product, service or company.

We provide consumers ' opinions to those who seek them for purposes of improving quality of service or products. Thus, the FKC is a way for consumers to communicate indirectly with the company they are dissatisfied about, and hopefully lead to an improvement in the situation. This is indeed the business model of the FKC, which makes money by selling access to valuable consumer opinion data to interested companies. To realize this concept, the FKC has been collecting user opinions since March 2015.

Consumers must register on the FKC service via its mobile application (iOS and Android) or its website. The registration form is a simple and can be filled by anyone who is capable of reading Japanese at a basic level. Figure 2 shows main functions in the FKC service. Users of the FKC can post their negative opinions from simple page (Right), also they can watch posts coming from other FKC users (Left).

The FKC rewards users with points in return for their posts. Once registered, users can post their opinions. Table 1 shows the schema of the post in our corpus. All the metadata fields are optional in order to simplify the post process as much as possible.

Point value grows with the opinion's quality (the length of the post, and other criteria). The point also increases as a user adds metadata relevant to the post (adding category, product/service name, company name). Table 2 shows the schema for the user profile. Most of the user profile in-

[2]http://rit.rakuten.co.jp/opendata.html

[3]http://mpqa.cs.pitt.edu/

Table 1: Contents to be posted as fuman

Field	Essentiality	Data type	Example (English translation)
fuman	mandatory	free text	電車が毎日、遅延してばかり (Train is behind schedule everyday)
proposed idea for fuman	optional	free text	余裕をもったダイヤにした方がいい。 (Train company should adjust a timetable)
target of fuman	optional	free text	東京線 (TokyoLine)
service provider of target	optional	free text	東京鉄道 (TokyoRailway)
sub-industry	optional	categorical	駅・電車 (Station & Train)
industry	optional	categorical	公共・環境 (Public Service)

Table 2: User profile

Field	Essentiality	Data type	Example (English translation)
gender	optional	categorical	男性 (male)
birth year	optional	integer	1990
job	optional	categorical	会社員 (employee)
state	optional	categorical	東京 (Tokyo)

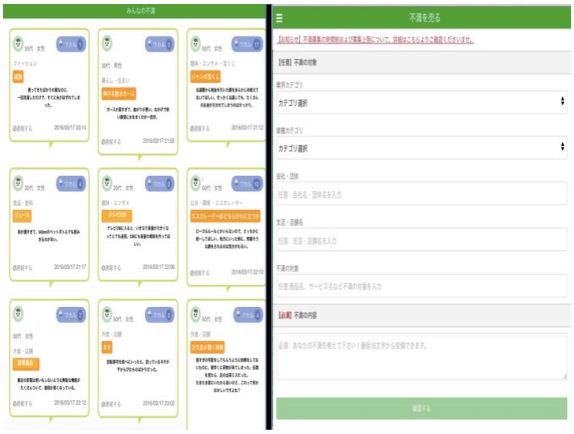

Figure 2: Main function of the FKC service. FKC users can post their opinions with a page for posting (Right), they can watch posts from other FKC users with time-line (Left)

formation is also optional to ease the registration process[4]. Thus, a post containing only a short sentence and with no additional optional fields set has the lowest value. The maximum price can only be reached by a quality post with all optional fields filled-in for a user who has filled in all their own personal information. This system promotes user to fill user profile.

3.1. Point and procedures for exchanging

As we mentioned above, the point in return of their posts has real monetary value, which is exchangeable with gift cards for an electronic commerce service. Mostly, this point is from 1 to 10 for a post, about 5 on average. As of March 2016, 1 point is always equal with 1 Japanese Yen. In Japan, a bottle of mineral water or a can of coke is around 100 Yen. Thus, around 20 posts have almost same value of a soft drink.

As of March 2016, the FKC is providing only *Amazon.co.jp gift card*® as an exchangeable gift card. FKC users are able to ask the FKC to exchange their points with the gift cards. There are 2 advantages to use the Amazon gift card. First, Amazon.co.jp is one of the most popular electronic commerce service in Japan, therefore, FKC users are able to purchase everything with the gift cards they get. Second, FKC users can receive a code of gift cards by e-mail, which makes procedures easy. The FKC sets 500 Japanese Yen as the minimum value of exchange, therefore, FKC users must accumulate at least 500 points to ask to exchange.

All exchange procedures are done via Internet. First, FKC user sends a request to exchange through applications of the FKC service. Second, the FKC reduces the point from accumulated user's point, and send a gift code of Amazon.co.jp gift card with e-mail. Finally, the user is able to use it.

3.2. Metadata in post and user profile

3.2.1. Metadata of post

The first metadata field is the *industry* and *sub-industry* of the company that the user post is about. Sub-industry is a sub category within the broader industry category. Table 1 shows an example whose industry is "Public Service" and sub-industry is "Station & Train". We have 14 industry categories and 10-13 sub-industry for each.

There is also a company/organization field, and a product/service name field that the user can either select from the existing list, or enter if there is not in our database yet.

3.2.2. Metadata of user profile

FKC users can register following 4 user profile information. The *None* are recorded in fields if a user does not choose.

Gender The gender is a categorical value which can be either male, female.

Prefecture (State) The area of residence, as a categorical value which can be set to any one of the 47 prefectures of Japan.

[4]Putting user profile is mandatory from December 2015 to collect more precise opinions and to know sender of opinions more precisely.

Birth year The birth year is a 4 digit integer.

Occupation (Job) The main occupation of the user. We set 12 typical occupations in Japan.

- 経営者・役員 [owner/board member]
- 会社員（事務系）[employee (office worker)]
- 会社員（技術系）[employee (engineer)]
- 会社員（その他）[employee (else)]
- 専業主婦（主夫）[housemaker]
- 専業主婦（主夫）[housemaker]
- 学生 [student]
- 公務員 [public employee]
- 無職 [no job]
- パート・アルバイト [part time]
- 自営業 [self-employed]
- その他 [else]

3.3. Annotation

An important feature of the FKC is that anyone can register and post anything as the content of their posts. Therefore, there will inevitably be some undesirable posts. To cope with such posts, native Japanese speaking operators manually annotate posts. They carry out three kinds of annotations; 1: label posts with a content-check flag, 2: correct category mistakes, 3: normalize the company and product name fields. For 2 and 3, we save *None* if a users does not choose them.

3.3.1. Filtering out unsuitable posts

All posts are assigned "content-check flag" label, which identifies a post as good or bad. A good post means it can get points, whereas a bad post should not result in any point reward. Since points given for posts have real cash value, there is a real business benefit for filtering out posts that are gibberish, incomplete/meaningless, uninformative or that use offensive words. Mainly, we give bad content-check flags by following reasons,

Duplication The post is completely same or extremely similar to already posted one.

No meaning sentence The post which has no meaning as Japanese is given bad flag. For example, "あああああああ (aaaaaaa)".

Positive opinion We give bad flag when a post means positive opinion, and has no negative opinion at all. For example, "きのう食べたカレーはとても美味しかった。あしたも食べたい！(It was delicious curry I ate yesterday. I would like to eat tomorrow!)"

Offensive Posts containing personal information or those that are offensive are marked as bad including those mentioning untitled civilians or containing racial discrimination or abusive words[5].
For example, "スーパー店員の山田太郎という店員の感じが悪かった。(A shopkeeper named Taro Yamada, he was disgusting.)"[6].

3.3.2. Correcting mis-categorized posts

It can unfortunately happen that users do not select the correct industry/sub-industry category given the content of their post. The operators check these categories and correct mistakes when they are found.
For example, "Public Service" is correct category in an example of Table 1. But some users might post their opinions as "Sightseeing & Leisure" if their posts' context is like "Negative opinions for passengers in a train when I was on the way to leisure places". In such case, the operators correct "Sightseeing & Leisure" to "Public Service".

3.3.3. Normalization of free-text fields

The company and product name fields allow direct user input. Since users can enter the same entity in multiple forms, this field must be normalized. For example, a user might mention "*Apple Inc.*" as "*apple computer*" or "vendor of iphone", which neither is the actual name of the company. To cope with such ambiguity problem, our operators manually normalize the data to an agreed upon single value, "*Apple Inc.*" in this case.
For now, we are carrying out normalization only for the "company/organization" field because "product/service name" field has such a sheer variety of products and services that users may be referring to that it is hard for operators to cope with all of them. The procedure for normalization is following:

1. We made a master database of representative manufacturing and hospitality companies in Japan. This is because most users mentioned about company.
2. We make relationship between master data and values in "company/organization" that user mentioned. If the master data does not have "company/organization", we clean up the text and add it into the master.

3.3.4. Annotation procedure

For annotation, we hired 8-10 part-time workers as annotation operators. Each post is annotated by only one worker. We put a priority on speed. The FKC is running platform and new posts are continuously being created[7]. Point reward must be done with a reasonably short delay for best customer service.
Given that each post is reviewed by only one part-time worker, we asked one of our employee, also a native

[5]We removed posts which are categorized into Offensive from the FKC corpus because this category includes sensitive contents.

[6]This sentence is just a fictional example. *Taro Yamada* is a common fictional name in Japan as same as *John Smith* or *John Doe* in American culture.

[7]As of September 2015, the FKC gets an average of close to 10,000 posts a day.

Japanese speaker, to double-check the annotations of the part-time workers. As this employee knows our rules well, we believe this system is enough to ensure the overall quality and accuracy of the annotations.

To reduce mis-annotations as much as possible, we run training sessions and our employee provides feedback. During the training phase, our employee explains annotation rules to part-time workers who then apply the annotation procedures to 1,000 posts. When they finish their annotation tasks, our employee checks mis-annotated posts and lets them know their mistakes in detail as feedback. Finally, we ask them to annotate incoming posts. Even if after training, our employee gives feedbacks to them if they make mistakes.

We recruited them in some ways; from SNS like Twitter or Facebook, introductions from our employees' families or friends. Some part-time workers live near from our office, others far from our office. Considering this situation, we asked them to work at their home for the purpose of making our procedure in uniform. Thus, all training and feedbacks are carried out with *Skype*® which is online conversation tool.

As a result of this training and feedback efforts, we have high agreement rate on "content-check flag" and correction of mis-categorized posts between our employee and part-time workers. We have 99.5% averaged agreement rate on all part-time workers for "content-check flag". And we have 99.2% on mis-categorized "industry" category and 99.0% on mis-categorized "sub-industry".

3.4. Data format

Our corpus is provided with JSON format data as shown in the upper part of Figure 3. The JSON format is easily converted into XML format because we put a script with the corpus. In this data, every item has post-meta-data and user-meta-data. The file size is around 180 MB with JSON format.

3.5. Corpus License

The FKC corpus is now available under an original license of the FKC, and is only for research purpose. Currently, our license is available only in Japan. We are working to make the FKC corpus available also for researchers in overseas.

To use the FKC corpus, the one have to make a contract with the FKC directly. The First author is helping to give permission to researchers. Those who hopes to use the FKC corpus needs to send e-mail to first author and ask permission to use.

4. Corpus statistics per device and user demographics

In our corpus, there are 254,683 posts and 25,092 users.

Table 3[8] shows some basic statistics about the devices were used to write posts on the FKC by its users[9]. The "others" category includes minor mobile devices as well as unknown devices. Most posts are from Android and iPhone mobile devices, with an almost 80% share of posts. The average character length of posts made on these two mobile platforms is 46-47 characters. Compared with posts from PC,


```
[
    {
        "normalized_company_name": "市役所",
        "product_category": "地方行政",
        "user_number": 1,
        "fuman": "収入がゼロでも徴収される糞制度。",
        "state": null,
        "product_name": "国民健康保険",
        "birth_year": null,
        "status": "ANNOTATED",
        "company": "市役所",
        "job": null,
        "gender": null,
        "industry": "政治・行政",
        "proposals": null,
        "time": "2015-03-18 22:35:42"
    },
    {
        "normalized_company_name": null,
```

```
<?xml version="1.0" encoding="UTF-8" ?>
<0>
  <normalized_company_name>市役所</normalized_company_name>
  <product_category>地方行政</product_category>
  <user_number>1</user_number>
  <fuman>収入がゼロでも徴収される糞制度。</fuman>
  <state /><product_name>国民健康保険</product_name>
  <birth_year /><status>ANNOTATED</status>
  <company>市役所</company>
  <job /><gender /><industry>政治・行政</industry>
  <proposals /><time>2015-03-18 22:35:42</time>
</0>
<1>
  <normalized_company_name /><product_category>その他</
```


Figure 3: Data format example of JSON (Up) and XML (Down)

Table 3: Statistics per device

device	#post	Avg.tokens	Avg.character
Android	102,378	26.867	46.734
iPhone	97,081	27.298	47.563
PC	48,372	34.404	59.346
iPad	5,436	20.788	50.075
others	1,416	27.207	50.995
total	254,683	28.608	49.692

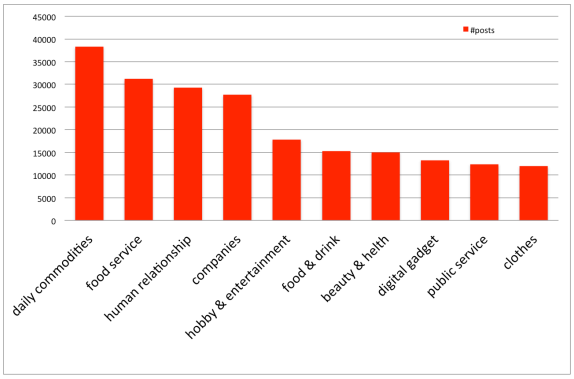

Figure 4: Top10 industry ranking. x axis is category name in industry attribute, y axis is #post for it.

Table 4: Statistics of content-check flag

content-check flag	#post	ratio
Good	241,678	0.948
Bad	13,005	0.052

[8] Statistics collected from tokenized posts using MeCab 0.996.

[9] Information parsed from User-Agent string

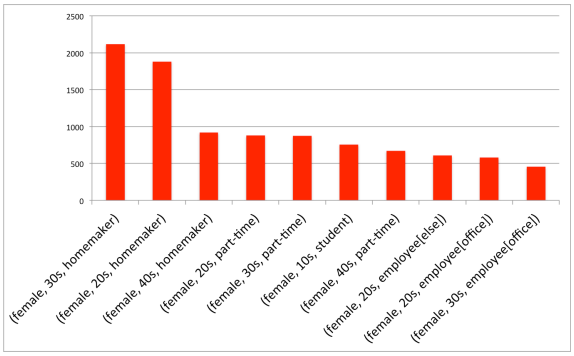

Figure 5: Top10 user demographic of (gender, age, job). x axis is (gender, age, job) and y axis is #user for it.

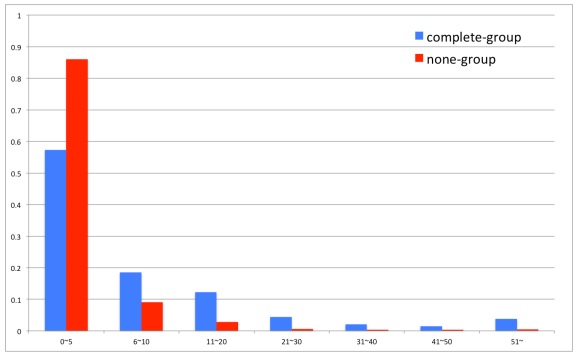

Figure 6: Distribution of #post for complete-group and none-group. x axis is segment of #post, y axis is ratio per group

posts from mobile devices are shorter about 10 averaged character length as well as averaged token length. It is assumed that users using mobile devices tend to express their opinions more briefly than PC users. According to Neubig and Duh (2013), the average character length among Japanese Tweets is 40-45. From this observations, we can say that posts in the FKC mostly came from mobile devices, and the its post length is close to Twitter.

Annotation operators have labeled the posts as good or bad and this information is stored in the "content-check flag" as in Section 3.3.1. Good posts far outnumber bad ones, by a ratio of almost 19 to 1, as 241,678 posts (about 95% of all) are good posts, and only 13,005 (about 5% of all) posts are bad. From this observation, we can say that the vast majority of FKC users follow the guidelines about writing good posts.

As for the "industry" and "sub-industry" fields in Section 3.2.1., 99% of posts have a "industry" category, and 96% have both of "industry" and "sub-industry". Figure 4 shows top 10 for posted industry categories. The top 3 categories count for as much as 39% of all posts. Considering the target of FKC users are ordinary Japanese consumers, such a selection of categories make sense, as they are such a common part of everyday life experience.

Figure 5 shows the top 10 for user demographics for the combination of gender, age and job. These top 10 combinations occupy about 38% of all users. This is a zipfian-like distribution where a few combinations are very common, followed by a long tail of all the remaining possible combinations. The post "industry" distribution for this top 10 group of user segments is almost entirely about "daily commodities", "human relationships" and "food service industry", mirroring the distribution of the whole dataset, meaning they are a good representative sample of all users of the FKC.

5. Correlation between user-meta-data and users' motivation to FKC

To make clear how a system of the FKC point incentive works on users' motivation for posting their opinions, this is shown by the relationship between tendency of filling in user-meta-data and users' posts. The FKC service lets FKC users know the point incentive system of the FKC when new users start to use FKC services. FKC users know that points grow up when they post their opinions, so that it is considered that motivated users fill in all user-meta-data and they post much with abundant post-meta-data.

In the FKC corpus, there are 3 types of users in the point of profile information. We call users who filled all user-meta-data as *complete-group*, users who do not put any of user-meta-data as *incomplete-group*, and users who have no user-meta-data at all as *none-group*. In the corpus, 66% of users is complete-group, 20% is incomplete-group, 14% is none-group.

We investigate the correlation with the number of post, filled ratio of post-meta-data, persistency ratio of posts. For this investigation, we omit incomplete-group because it is considered that users in incomplete-group understand the FKC incentive system, however, they still refuse to put all their user profile information by any reasons. Thus, we compare complete-group with none-group in 3 investigations. In all of 3 investigations, we observe positive tendency for the FKC service.

5.1. Correlation with the number of post

If users in complete-group are motivated by the FKC point incentive system, they post much than users in none-group do. On average, users in complete-group prove to be much more prolific than users in none-group. In fact, complete-group users post an average of 11.99 posts compared with none-group users who only post an average of 3.51 posts each. Figure 6 presents distribution of #post for complete-group and none-group. We observe that the post ratio of complete-group is high in segments of much posts (all segments in more than 6 posts) compared with none-group. The ratio of users who post more than 50 posts is 3% in complete-group, by contrast, 0.4% in none-group.

5.2. Correlation with persistency ratio of posts

We show correlation between user-meta-data and the number of post in Section 5.1, however, there is possibility that some users are just new to the FKC and their posts are still a few. Considering this possibility, we might not say correct correlation from the ratio. Thus, we investigate users' continuity of posts. If the FKC point incentive works as motivation to users, it is presumed that they keep posting their opinions to the FKC to accumulate the FKC point. It

Figure 7: Persistency ratio of complete-group and none-group. x axis is *k* (days after first post) and y axis is persistency ratio

is desirable for the FKC that users keep posting their opinions because we are able to construct users' models based on users' demographic information if the FKC has enough posts coming from each user.

We define *persistency ratio* to present how much FKC user keep posting their opinions after their first posts. Here, for a user $u \in U$, we call the first date when *u* posted first opinion as *t_u_0*. With t_u_k, we count the ratio which *u* posted in the t_u_k day. Still, there is possibility that *u* did not post his opinion in the just t_u_k day. So, we use an adjustment parameter α to denote before and behind t_u_k day. With the α parameter, we can check whether *u* posted his opinion in the range $range\_t\_u\_k$: $[t\_u\_k - \alpha, t\_u\_k + \alpha]$ or not. The persistency ratio is defined with the following formula.

$$Persistency\ ratio = \frac{\Sigma_{u \in U} count\ post(u, k, \alpha)}{\mid U \mid}$$

$$count\ post(u, k, \alpha) = \begin{cases} 1 & if\ u\ post\ in\ range\_t\_u\_k \\ 0 & else \end{cases}$$

where

- $range\_t\_u\_k : [t\_u\_k - \alpha,\ t\_u\_k + \alpha]$
- $\mid U \mid$: the number of users

Figure 7 shows the persistency ratio when *k* of t_u_k is 2, 5, 8, 15. We use $\alpha = 1$ when *k* is from 2 to 8, $\alpha = 2$ when *k* is 15. As shown in the Figure 7, even though the difference in persistency ratio between complete-group and none-group is shrinking as *k* increases, there is always 2-3 times difference. From this tendency, there is clear correlation between persistency ratio and user-meta-data. We can conjecture that users in complete-group tend to be well motivated with the FKC point incentive system, therefore, they keep positing than none-group which is less motivated group.

5.3. Correlation with filled-in ratio of post-meta-data

If users in complete-group are motivated by the FKC point incentive system, we can assume that they put more post-meta-data to get more points. In other words, users in complete-group tend to have less *None* value in their posts.

Table 5: Distributions of #None in post-meta-data

#None in post-meta-data	complete-group	none-group
0	19%	17.8%
1	25.9%	23.2%
2	29.9%	26.1%
3	24.7%	29.4%
4	0.5%	2.5%

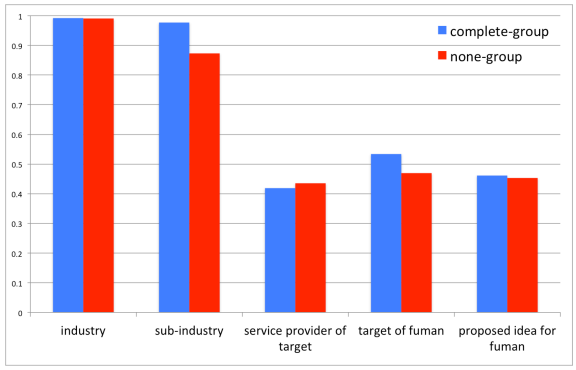

Figure 8: Input ratio of post-meta-data. x axis is attribute of post-meta-data, y axis is input ratio of it

Table 5 shows distributions of #None in post-meta-data. The complete-group has much ratio when #None in post-meta-data is from 0 to 2 compared with none-group. Interestingly, the most highest ratio in complete-group is when #None in post-meta-data is 2, by contract, the one in none-group is when #None in post-meta-data is 3. This means there is tendency that users in complete-group put plus one metadata than users in none-group.

Figure 8 shows input ratio of post-meta-data. For both of complete-group and none-group, there is a tendency that "service provider of target", "target of fuman" and "proposed idea for fuman" are small ratio than others. These meta information are detailed information. Therefore, it is presumed that users do not remember such detailed information and skipped filling in.

The difference between complete-group and none-group is mainly in "sub-industry" (10% difference) and "target of fuman" (6.5% difference). We infer that users' proficiency level relates to this difference. Considering 10 - 15 "sub-industry" categories per one "industry" category, users need to comprehend category structures to fill in. Also for "target of fuman", users are required to remember product or service names to fill in. Even though users need to understand well to fill in these post-meta-data, we suppose that the FKC incentive system works as motivation for filling in.

6. Applications for NLP tasks

Many NLP tasks can make good use of the FKC corpus. The abundance of user profiles in the FKC corpus makes it especially suited to the author inferring task. One example is Nguyen et al. (2011), where they use blog corpus to construct models with the objective of predicting the author's age. Mukherjee and Liu (2010) targets gender prediction, also from a corpus sourced from blogs. The FKC corpus can be a useful corpus to support both of these targets, as its

user profiles include both age and gender information. The FKC corpus has also other features, such as users' state, job and posts' industry categories, which are high-potential effective features.

Domain Adaptation is a task which trains a model on labeled-corpus to predict labels for other unlabeled-corpus. Dai et al. (2007) proposed domain adaptation metric between similar dataset, and Xiao et al. (2013) proposed a model between not-very similar documents such as news text and product reviews. The FKC corpus is useful again as a labeled training data for such domain adaptation models because the corpus has industry and sub-industry category for almost all posts, and there are various industry categories as in Figure 4.

7. Conclusion amd Future work

In this paper, we have presented a new corpus that is consisted with lively coming negative opinions. This corpus is useful for various kinds of NLP research and we have presented some NLP metrics in which our corpus is applicable. This corpus is useful in following point: First, all posts are from ordinary consumers, which is valid datasource of opinion mining. Second, this corpus has rich metadata, which is essential information for supervised machine learning methods. Third, this corpus is less noisy compared with existing datasets of SNS because the corpus contains only negative opinions.

We showed some correlations between an inventive system of the FKC and users' motivation to keep posting their opinions with much metadata. Even though we observe positive tendency between user-meta-data and users' motivation, however, it is hard to assert causal relation clearly. We are not able to investigate how the FKC point incentive system (point incentive from user profile) works on users' behaviors because the FKC does not save all log that users changed their user-meta-data in FKC service. Besides, it is hard to conduct this analysis with the current FKC service because putting user-meta-data is mandatory from December 2015 to collect more precise opinion and to know sender of opinions more precisely. Therefore, we are planning to investigate users' behaviors via questionnaire survey, like "how do they feel about the FKC incentive system?" or "have you ever tried any of questionnaire or survey service with incentive?"

In the near future, we will publish a new version with more posts. And we will extend data input method and metadata on it. We are currently working on a new system which accepts post without registration. With this system, new posts from wide variety of users will be increased. And we believe that new metadata will lead to new applications of machine learning methods.

8. References

Boley, D., Gini, M., Gross, R., Han, E.-H. S., Hastings, K., Karypis, G., Kumar, V., Mobasher, B., and Moore, J. (1999). Partitioning-Based Clustering for Web Document Categorization. *Journal of Decision Support Systems*, 27(3):329–341.

Dai, W., Xue, G.-R., Yang, Q., and Yu, Y. (2007). Co-clustering Based Classification for Out-of-domain Documents. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 210–219.

Mukherjee, A. and Liu, B. (2010). Improving Gender Classification of Blog Authors. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing(EMNLP)*, pages 207–217.

Neubig, G. and Duh, K. (2013). How Much is Said in a Tweet? A Multilingual, Information-Theoretic Perspective. In *AAAI Spring Symposium on Analyzing Microtext*, pages 32–39.

Nguyen, D., Smith, N. A., and Rosé, C. P. (2011). Author Age Prediction from Text Using Linear Regression. In *Proceeding of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH'11)*, pages 115–123.

Noll, M. G. and Meinel, C. (2008). Exploring Social Annotations for Web Document Classification. In *Proceedings of ACM Symposium on Applied Computing (SAC)*, pages 2315–2320.

Orimaye, S. O., Alhashmi, S. M., and Siew, E.-G. (2012). Natural Language Opinion Search on Blogs. In *Proceedings of 12th Pacific Rim International Conference on Artificial Intelligence (PRICAI)*, pages 372–385.

Pak, A. and Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, pages 17–23.

Sasa, P., Miles, O., and Lavrenko, V. (2010). The Edinburgh Twitter corpus. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics in a World of Social Media*, pages 25–26.

Schenker, A. (2003). *Graph-theoretic Techniques for Web Content Mining*. Ph.D. thesis, University of South Florida, Tampa, FL, USA. AAI3182715.

Uryupina, O., Plank, B., Severyn, A., Rotondi, A., and Moschitti, A. (2014). Sentube: A corpus for sentiment analysis on youtube social media. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC)*, pages 4244–4249.

Wiebe, J. and Theresa Wilson, C. C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3):165–210.

Xiao, M., Zhao, F., and Guo, Y. (2013). Learning Latent Word Representations for Domain Adaptation using Supervised Word Clustering. In *Proceedings of the Conference on Empirical Methods on Natural Language Processing (EMNLP)*, pages 152–162.

Zhang, Z., Wu, G., and Lan, M. (2015). ECNU: Multi-level Sentiment Analysis on Twitter Using Traditional Linguistic Features and Word Embedding Features. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval)*, pages 561–567.