Construction of a Multilingual Annotated Corpus for Deep Sentiment Understanding in Social Media

YUJIE LU[†1] KOTARO SAKAMOTO[†1] HIDEYUKI SHIBUKI[†1]
TATSUNORI MORI[†1]

The surge of social media makes it possible to understand people's emotion in different cultures. In this paper, we construct an annotated corpus for multilingual sentiment understanding. The annotation is developed in a multilingual setting including English/Japanese/Chinese, and on a representative dataset including 4 topics (spanning 3 genres, which are product, people, and event).To deep understand expression mechanism of feeling entailed in the text, we labelled sentimental signal words and rhetoric phenomenon in addition to overall polarity. This innovative corpus can be a helpful resource for the improvement of sentiment classification, cross-cultural comparison etc.

1. Introduction

With the surge of social media websites such as Facebook, Twitter etc., a huge amount of user-generated content has been created out from their prevalence. Users post their opinions on a variety of targets, such as commodity, people and events in real time. Based on these messages, companies can investigate consumer reaction to their products; political parties can understand the popularity of their candidates among the voters to forecast the election, and public opinion to a social event can be revealed faster than traditional polls. These facts bring researchers an unprecedented chance to leverage them for the purpose of scientific study and many useful applications have been put forward so far (Liu, 2012).

However, although there have been some progress, key challenges still remain. Firstly, the dynamic nature of social media with subtle ways of expression has implications for the study of sentiment analysis. Many work(Preslav Nakov,2013 ; C´ıcero,2014 ; Bing Xiang,2014) carried out on SemEval dataset showed that highest accuracy of the sentiment analysis on social media is around 70% ,which is not high comparing with the same work on other texts(88.3% on IMDB Dataset,93.7% on Polarity Dataset (HuiFeng Tang,2009)).

Secondly, although social media makes it possible to offer people collections of multilingual messages, little work has been done on culture diversity or regional comparison. Alexandra Balahur (2013) discussed the implementation of sentimental analysis on multiple languages by machine translation. Svitlana Volkova (2013) showed how the gender information impact the effect of polarity classification in different languages. However, these works didn't observe culture difference regarding same evaluation objects among different regions.

To tackle the first challenge, we need understand the key difference between tweet text and general text, such as newswire, review, etc. Tweets are expressed in a more flexible, casual way and the sentiment contained in them are usually subtle and underlying. Take the following tweet as an example:

Wow, with #iPhone6, you can send a message just by talking! In any voice you like. So can my mom's old rotary dial.

First, in this example, we see special symbols commonly used in social media like # (a topic). # and other symbols show the specificity of superficial expression form in social media. The difficulty cause by these symbol can be alleviated by a combination of preprocessing. In terms of deep level of emotional understanding, we find it is not as easily as it looks. In the first sentence, the author is praising Iphone6, but the second sentence turn to be a criticism by comparing it with something 'old'. As a whole, this is a sarcastic tweet which strengthens the looking down on IPhone 6.For tradition methods heavily depending on sentiment dictionary will probably output 'positive' since there are more positive words. It's hard for system to tell the rhetorical devices which human beings can percept.

Therefore, to better understand the flexible expression ways of social media users, it's necessary to observe the real tweets and reveal the patterns within them by human beings. If the clues or hints for showing the sentiment of a message can be annotated as the following way, system can learn underlying patterns from a certain amount of data points having the same language phenomenon.

Wow [positive], **with #iPhone6, you can** [positive] **send a message just** [intensifier] **by talking! In any** [intensifier] **voice you like** [positive]. **So can my mom's old** [negative] **rotary dial** [comparatively neutral]. [sarcastically negative]

In fact, if we look into the failure cases of the sentiment classification, we find those sophisticated tweets often contains rhetoric phenomenon. Rhetoric is one of the reasons for the fallback of sentiment analysis (Xie Lixing, 2011; Michael Wiegand, 2010). It is difficult to decide the polarity of example tweet based only on the sentiment dictionary and surrounding features around those sentimental words without mine out the comparison and the sarcasm. Rhetoric phenomenon is needed to be discussed in depth in social media.

For multilingual setting, concerning the linguistic difference, a

†1 Graduate School of Environment and Information Science, Yokohama National University

ⓒ2015 Information Processing Society of Japan

multilingual golden dataset for social media are highly expected. Although there have been some annotated datasets for social media, they are proprietary, only in one language or tagged at a shadow level (usually only global polarity). To our best knowledge, regarding distant multi-languages such as Chinese, English and Japanese, there isn't an existed annotated dataset for social media. Therefore, to tackle the abovementioned two challenges, it is meaningful to construct such dataset to help the research on sentiment analysis in multilingual setting in social media.

This paper introduce our work on a construction of multilingual annotated corpus for deep sentiment understanding in social media. Our contribution and characteristics are as following:

Firstly, we conduct deep annotation on the dataset including 4 topics (spanning 3 genres, product, people, and event). Apart from overall polarity, annotators are asked to tag the rhetoric phenomenon in a tweet, emotional words having influence on the overall polarity, degree words around those emotional signal and sub-topics in a tweet.

Secondly, unlike the existed multilingual language work based on close languages, our three languages(English, Japanese, Chinese) are distant from each other, which help us to magnify the variation between languages.

Thirdly, we carry out a Pivot Data method to help us improve the agreement between annotators. The Kappa Statistic shows this method works well. This method especially suits for the non-factual annotation.

The structure of the rest of the paper is organized as follows: Related work is presented in Section 2.Section 3 mainly introduce the data collection and the data selection for annotation. Section 4 shows the process of the annotation including its preparation and management. Section 5 is the analysis of the annotation result, making comparison between different languages and topics. Section 6 is the agreement analysis and the deficiencies. Lastly, we draw the conclusion and put forward future work in Section 7.

2. Related Work

To evaluate and improve proposed methodologies, annotation is often applied to a dataset (usually a small portion of the whole corpus) expecting to find out underlying patterns contained in text. In the same way, for the different purposes of the researchers, there have been some existed standard annotated datasets in the field of sentiment analysis. We discuss them in this section and show their difference from our annotation.

2.1 Traditional Dataset for Sentiment Analysis

Movie review, newswire and product review etc. are traditional research objects for sentiment analysis (also called as opinion mining).Bo pang (2002) used Movie Review Data to testify the effectiveness of machine learning methods for sentiment analysis. The newest version consists of 1000 positive and 1000 negative processed reviews. Wiebe et al (2005) made a dataset called MPQA which contains news articles from a wide variety of news sources with many other states like beliefs, emotions, sentiments, speculations etc.

Bing Liu (2005) gathered thousands of consumer opinions from online customer review sites (called Pros and Cons Dataset), and discussed a new technique to identify product features. Murthy Ganapathibhotla (2008) collected hundreds of comparative sentences from product review websites and online forums, studied the method of mining opinions from them though identifying preferred entities. Miyazaki et al (2010) proposed a model separating the description of and opinion to products, discussed the method of decreasing in accordance among annotators and annotated a collection product reviews from online commercial site for helping improve the information extraction. Unlike these dataset comprising long and normal text, our text from social media only consist of no more than 140 characters and their way of expression is unstable, all of which brings difficulty to annotation even for human being.

2.2 Twitter Dataset for Sentiment Analysis

As to social media, SemEval 2013 Task 2(also 2014) (Preslav Nakov, 2013; Sara Rosenthal, 2014) offers two kinds of datasets — Subtask A tagged the polarity of the marked instance in the tweet; Subtask B tagged the overall polarity for nearly 10 thousand English tweets (together in two years). Many related researches (Alexandra Balahur, 2013; Bing Xiang, 2014) used Task B as their experiment data for different purposes. Spanish TASS corpus (Villena Rom án, 2013) is a Twitter corpus for Spanish which consists of 7219 Twitter messages tagged with global polarity and entity polarity if there is. SIEVE corpus[a] and sanders corpus[b] are two available dataset offered by private companies. Other than the former is proprietary, the latter can be accessed freely which contains thousands of tweet tagged with global polarity.

In terms of rhetoric aspect, Yi-jie Tang (2014) build a Chinese irony microblog data set containing 1000 messages, which he claimed to be the first irony dataset for Chinese. Gonz ález-Ib áñez (2011) build an English Twitter dataset by using the #sarcasm# hashtag and compared the performance of machine learning techniques and human judges on the sentiment classification. Although these works paid attention to the rhetoric phenomenon in the social media, most of them mainly concerned on one kind of them. However, our dataset involves more rhetoric types at the same time and we are further able to observe this slight difference of the use of rhetoric between cultures based on the multilingual setting.

2.3 Multilingual Dataset for Sentiment Analysis

For multilingual using, Svitlanna Volkova (2013) constructed a Tweet dataset including English, Spanish and Russian by Amazon Mechanical Turk, compared the variation of gender information in the three different languages, and showed that gender differences can effectively be used to improve sentiment analysis. Zornitsa Kozareva (2013) collected and manually annotated a metaphor-rich texts with the polarity and valence scores for four language including English, Spanish, Russian and Farsi, showed that the proposed technology for polarity and valence prediction of metaphor-rich texts is portable and works well for different languages. Alexandra Balahur(2013) constructed a multilingual dataset by translating English Tweet to Italian, Spanish, French and German, tested the performance of

a http://www.i-sieve.com/

b http://www.sananalytics.com/

ⓒ2015 Information Processing Society of Japan

the sentiment analysis classifiers for the different languages concerned and showed that the joint use of training data from multiple languages is effective. The languages used in the corpus constructed by these multilingual research are close (all of them belong to Indo-European language family), while our three languages are very distant (they belong to three different language families.).Besides, our level of annotation is much more detailed than these corpus which basically only have overall polarity, which guarantees the possibility of observing the word-level features inside the context.

Finally, our multilingual corpus was constructed on the same objects. This allows us to observe people's opinion in different regions from a macro perspective, which ensures the following research on observing public opinion and sub-topics people concern in different cultures from the annotation result analysis.

3. Data Collection

3.1 Evaluation Object Selection

From the view of cross-culture study (macro perspective), we not only want to build a corpus with fine-grained tags, but also we expect to build one that can be supportive to the future system development that can visualize the sentiment distribution information, display precise emotion evolution trend and illustrate comparison on sub-topics people care about in different regions or cultures. To fulfil this concept that unveil the differences, the very first step is looking for common and controversial topics discussed among these languages. In our research, we employed 6 international topics in 3 genres (product, people, event), namely Iphone6, Windows8, Obama, Putin, Scotland Independence and Japanese whaling as our evaluation objects. The query words are listed in Table 1.

Table 1: The query keywords for data collecting

Object	Code	English	Japanese	Chinese
Iphone6	I6	#Iphone6 lang:en	#Iphone6 lang:ja	Iphone6
Windows8	W8	#Windows8 lang:en	#Windows8 lang:ja	Windows8
Obama	OB	#Obama lang:en	オバマ	奥巴马
Putin	PU	#Putin lang:en	プーチン	普金
Scotland Independence	SI	Scotland Independence lang:en	スコットランド独立	苏格兰独立
Japanese whaling	JW	Japan Whaling lang:en	捕鯨	日本 捕鲸

3.2 3.2 Data Collection

In this part, we discuss the data collecting methods. As to the source of our data, we collect the data from Twitter.com by Twitter RESTful API same as many other researchers for English and Japanese. Given to the low quality of messages in Chinese on Twitter, we decided to use data from Weibo.com, a well-known Chinese-version Twitter, as a substitute. To Twitter source, we automatically collect data by the implementation of REST Search API using Tweepy. To weibo source, because the service provider doesn't offer search API openly unfortunately, we resorted to a crawler fetching results directly from search.weibo.com.

The collecting starts from 2014.10.19 and still on-going. This time we use data from 2014.10.19 to 2015.05.18(7 months). Notice that Weibo.com limits the maximum number of search result pages for one day as 50, so we only fetch those original weibos to avoiding duplication. Table 3 shows the number of data we collected (total number and average number per day) and their basic statistics (the average number of reply, favorite and retweet (only Weibo) per message).

In consideration of the convenience of management and following use, we further transferred and stored the original texts into database. Since the return of Twitter API is in JSON type, it's easy to process them. For Weibo text that are contained in HTML file mixed with HTML tags. We designed extraction patterns based on HTML structure and extracted elements needed by HTML parser. The extracted elements are then stored into database.

3.3 Data selection

As shown in Table 3, the scale of the corpus is very large, which makes it impossible to annotate all messages of them. Therefore, selecting representative messages from this corpus for the next stage is desirable. Social media such as twitter contains many messages that are commercial, news, etc. Those objective messages are of low value in the annotation stage.

In our research, we design a two-stage method to choose messages for building up a balanced annotation dataset. For each topic in each language we annotate 450 tweets. In the first stage, we use objective patterns to veto the unsatisfactory tweets, which means if one tweet contained one of these patterns, it will be removed from the candidate set. Table 2 shows the examples of the veto patterns used in this stage.

Table 2: Patterns used for exclude objective tweets

Pattern example	Description
^rt	Pattern indicates the tweet is a retweet.
[a-zA-z]+://[^¥s]*	Pattern indicates the tweet contains URL.
【.+?】	Pattern indicates the tweet is a commercial in Japanese, or news in Chinese.
(J)限定|在庫|施策|特価 … (E)news |breaking| … (C)分享|资源|共享…	Word patterns indicate the tweet is objective (commercial, news, Q&A etc.) for different languages.

In the second stage, we do the selection in a more soft way. We rank the tweet by the number of the @symbol, #symbol and number it contains. This method bases on the hypnosis that if a tweet contains more non-language word, it is more like to be a subjective message. This threshold differs from languages and topics, usually we set it as 2-4.

After the filtering by each stages, we will select a set of tweets whose length is longer than a certain value. The second stage

ⓒ2015 Information Processing Society of Japan

Table 3: The statistics of the number of tweets (2014.10.19~2015.05.18)

	English				Japanese				Chinese				
Item	Total #	Day Avg.#	Reply Avg.#	Favorite Avg.#	Total #	Day Avg.#	Reply Avg.#	Favorite Avg.#	Total #	Day Avg.#	Retweet Avg.#	Reply Avg.#	Favorite Avg.#
I6	2010370	9482.9	2.2	243.8	906376	4275.4	1.6	182.5	173493	818.4	26.9	10.7	4.5
W8	93231	439.8	1.5	341.6	37230	175.6	1.5	10.1	16807	79.3	2.8	1.7	0.6
OB	3019685	14243.8	3.5	153.5	915845	4320.0	4.6	476.5	104105	491.1	22.2	9.0	35.2
PU	631254	2977.6	3.1	118.6	637605	3007.6	3.3	4896.0	121884	574.9	11.5	3.6	14.7
SI	105272	496.6	4.1	106.5	19914	93.9	2.9	647.8	3326	15.7	64.5	18.3	38.5
JW	40689	191.9	3.4	50.7	177948	839.4	3.0	1161.2	486	2.3	2332.2	843.3	2635.0
Total	5900501	27832.6	3.2	157.0	2694918	12711.9	3.2	1910.1	420101	1981.6	23.3	8.9	18.23

won't be carried out if the number of candidate set is not large. The choice of the length depending on the scope of the candidate set. If the candidate set is large, we can select more long tweets from them; if the candidate set is small, we will reduce the length of the length threshold. A general setting for length is 100.The filtering work help us delete a large portion of undesired tweet in the database, by which lesson the time and effort for picking up suitable tweets. For keeping the randomness of the selection and the diversity of the text to the greatest extent, we didn't manually interfere except for the last candidates to avoid useless messages that failed to be filtered out.

4. Corpus Construction

4.1 Annotation Setting

Given to the limitation of funding, we chose both the two product objects and picked one from each of the other two genres as our annotation objects. For each object, three different editors carried through the annotation on it independently according to a common rule set. For each language, there are six annotators. (Table 4 shows the distribution of the annotators (A1-A6)).

Concerning the sentiment in social media text can be subtle sometimes, each of the annotator is native speaker or has the same proficiency as native speaker. In specific, annotators for Japanese text are all native Japanese college students; annotators for Chinese text are all the Chinese native graduate students. In consideration of the geographical wide use of English, we chose to build a comprehensive group which consist of 2 American, 1 Australian, 1 Indian and 2 European.

Table 4: The distribution of annotators (one language)

Topic	No.	A.(1)	A.(2)	A.(3)	A.(4)	A.(5)	A.(6)
I6	450	☑		☑		☑	
W8	450		☑		☑		☑
PU	450	☑		☑		☑	
SI	450		☑		☑		☑

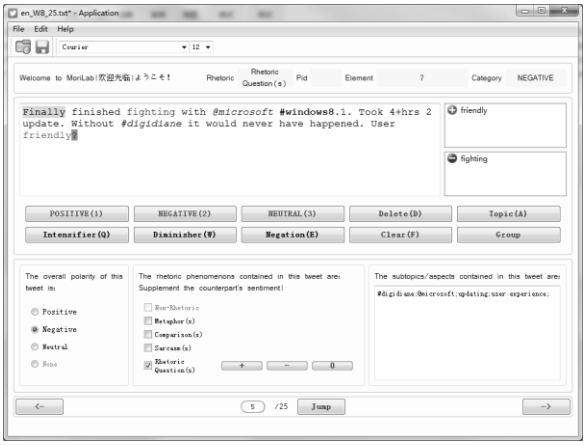

Figure 1:The interface of the annotation tool

Every annotator has 20 working hours to finish two topics for which he is responsible. To improve the speed and quality of the annotation work, a guidebook and an exercise beforehand were distributed or set before the real work. Moreover, in order to make it more convenient for annotators to tag information to words, we developed an annotation support tool. With the help of the tool, annotators almost don't need to input anything (except for the editing of sub-topics).Most of the tasks can be done by mouse and shortcuts. Figure 1 shows the interface of the annotation tool.

4.2 Annotation Tasks

In this section, firstly we inspect the way of emotion expression. Then, we introduce the annotation scheme, namely the detail of annotation standards. The result statistics and the annotation difference will be discussed in Section 5.

4.2.1 Way of Expressing feelings and emotions

According to the observation of tweets in the datasets, we found that there mainly are 2 general ways to express human being's emotion — the direct way and the indirect way. In addition, the indirect way primarily consists of four kind of rhetorical devices, which are sarcasm, comparison, rhetorical question and metaphor. There is a very significant point that although the three languages are so different at the word level or the syntax level, these basic sentiment expression ways are shared among them. As a result, we only give English examples here. Figure 2 shows the example tweets in different expressing ways. Although examples

ⓒ2015 Information Processing Society of Japan

Direct Way of Expression

(1) Negative: Getting really p*ssed off with #Windows8 it really is crap! The 'Search' facility doesn't work properly & now its lost some of my pics!
(2) Positive: It's beautiful. The resolution of pictures & videos, the screen size, the slow-motion capability when making videos & more. Amazing #iPhone6
(3) Neutral: At first I disliked #Windows8, but after using a #Surface I find it decent. But I think its best used as a tablet OS rather than a PC OS

Indirect Way of Expression

(4) Sarcasm: Every time I use #Windows8, I become more impressed with how profoundly bad a UX it is. Its an almost perfect #antidesign$
(5) Comparison: My 1983 #acornelectron started quicker than this hunk of junk #Dell xps #windows8 - Thank god for #macbookair
(6) Rhetorical question: What's the point of Westminster devolving powers to Scotland when the SNP will constantly demand complete independence?
(7) Metaphor: With each ramped up aggressive speech #Putin looks more and more like Hitler. His speeches against contrived enemies are identical

Figure 2: Tweets in different expressing ways

illustrating the idea in Figure 2 are simple, real tweets are much more flexible and can contains more than one expression ways simultaneously. The units we using for the tasks are word, sentence and message.

4.2.2 Emotional signals and their degree modifiers

As shown in Figure 2, words in the text are the smallest emotional unit (called as signal). In Example (2), words like 'beautiful' and 'amazing' are positive signals for evaluation object Iphone6. There are there types of emotional signals, which we define them as follows:

- Positive signal: words showing the good properties of the evaluation object;
- Negative signal: words showing the bad properties of the evaluation object;
- Neutral signal: words that imply a neutral context or user having not decided his opinion on the object.

(8)I really don't understand why anybody like WinRT application design. I think this is an amazing approach! #WinRT #Windows8 #NoMoreWPF

The great difference between a signal and an emotional word is a signal is where we can infer the global polarity while an emotional word is not necessary. In Example (8), even though 'amazing' is a positive word, it's not considered as a signal because it isn't related to the evaluation object Windows 8. A signal can be more than one word, such as phrases or idioms etc. In Example (1), phrase 'p*ssed off' are a negative signal.

Annotators mark the signals with either their dictionary emotion (in general) or their emotion in social network use. This means that an emotional word isn't vulnerable to the sentence polarity or the global polarity. In Example (4), although 'impressed' and 'perfect' are in the irony context, their emotion are always positive.

In addition, we define three types of degree modifiers for two reasons. Firstly, degree words are important surrounding information for the signals. Secondly, degree words can help annotators to distinguish the boundary the signal, which is crucial for none-space languages such as Chinese and Japanese. The types of degree modifiers are defined as:

- Intensifier: words that strengthen the emotion they modifies, such as 'very'/'really' etc.;
- Diminisher: words that weaken the emotion they modifies, such as 'a little'/'almost' etc.;
- Negation: words that reverse the emotion they modifies, such as 'not' etc.

(9) @Microsoft really? I updated Win 8.1 because it should fix problems, not generate more troubles!! #windows8 Sucks!!!

For each language, there is only a limited number (tens) of ways to express the extent, especially negation. Degree modifiers are usually explicit, such as 'really' in Example (1), 'almost' in Example (5) and 'doesn't' in Example(1).However, implicit negation is also allowed. In Example (9) both 'should' are negation to its modifying signal 'fix problems'.

Lastly, emotional signals and degree modifiers are usually appear in pair such as 'profoundly bad' in Example (4),'doesn't work' in Example (1) etc. Solely tagged degree modifiers are required to avoid.

4.2.4 Global Polarity

Global polarity is the fundamental information for the sentiment corpus, which is annotated by almost every sentiment classification-oriented corpus. In our task, the global polarity are divide into three categories. They are defined as follows:

- Supportive: messages that show author's good intention to evaluation object;
- Non-supportive: messages that show author's malicious intention to evaluation object;
- Neutral:
  - Messages that are truly neutral (half supportive and half non-supportive)
  - Non-comment messages, irrelevant messages, and objective messages etc.

ⓒ2015 Information Processing Society of Japan

(10) The nationalist criticism of the Smith Comission report is that it isn't independence. That's because Scotland didn't vote for that.

Example (1) (2) (3) are supportive, non-supportive and neutral message respectively. Example (8) are irrelevant message. Unlike the messages above-mentioned, some message can be blurry, of which global polarity are easy to be affected by the understanding or mentality of the annotators.

Global polarity of Example (10) can be either non-supportive or neutral depending on how do annotators understand the second half. It can be interpreted as a mocking of Scottish or a pure statement. For these kinds of tweet, a majority decision will be adopted to decide the global polarity when building the merged golden dataset.

4.2.5 Sub-topic Information

Sub topics are necessary to investigate the component structure of people's opinions of an evaluation object. Moreover, the inherent polarity of an emotional word may changes with the sub topics. For these reasons, the sub topic information are required to record.

(11) Just picked up an #iPhone6 the screen is beautiful, but my god is it large! Crossing my fingers it doesn't bend!

Annotators can firstly extract sub topics direct from the message (this also records the position of the sub topic) and then modified them into more standard shape. If there is no apparent sub topic words in the message, annotators can summarize the message. At least one sub topic is needed. For example the sub topics for Example (11) are screen, size and bending. Screen can be extract directly from the text, and size can be obtained by summarizing the next part. As to bending, it requires the observation of other messages. Not until we know there are a couple of message talking about the bending problem, it's difficult to pick up bending as a sub topic.

5. Annotation Result Statistics and Analysis

After the first round annotation work, we further endeavored to work a Pivot Dataset by observing and merging the results of the three annotated dataset on the principle of majority decision. Meanwhile, for those tweets that didn't get accordance in the first round, a double check with a fourth judge is carried out to guarantee the quality. Finally, based on the comparison with the pivot dataset, we ask all the annotators to fix their own annotation again by which we can correct the errors in the first round. Concerning individual differences on sentiment judgment, annotators decide whether to change or stay with their original global polarities at this step. After the last round annotator check, we will get the final golden dataset and checked annotator datasets.

Although all the original annotator datasets, the checked annotator datasets and the golden dataset are going to be distributed openly, the following statistics are conducted on the original annotator dataset unless otherwise specified. The original data is more suitable to observe the initial situation of the annotation result and can show us the tendency of the dataset. We will also observe this numbers of the final golden dataset in our future work.

5.1 Emotional signals and their degree modifiers

Table 5 showed the average number of signals and their modifiers per message. According to Table 5, we found that the Chinese users use the most emotional signals closely followed by English users. As expected, the neutral signals are much less than the other two in all the three languages while it's interesting to see that the number of positive signal and the number of the negative signal are very close. Take English 'Putin' as example, the two numbers are both 1.25, which means there are the same amount of emotional signal in the collection overall. However according to Table 7, the global polarity of English 'Putin' differs sharply. This imply the leverage of emotional signal alone is far from enough to decide the global polarity even at a collection level,

Table 5: Average Number of Signals and Their Modifiers per Message

	Average Number of signals and their modifiers Per Message						
Item	**Positive**	**Neutral**	**Negative**	**Intensifier**	**Diminisher**	**Negation**	**Sub-topics**
I6	1.23	0.04	0.89	0.97	0.13	0.25	2.51
W8	1.20	0.01	1.10	0.51	0.07	0.21	2.96
PU	1.25	0.04	1.25	0.48	0.06	0.30	3.02
SI	0.81	0.04	1.07	0.26	0.03	0.23	2.87
EN	1.12	0.03	1.08	**0.56**	**0.07**	0.25	2.84
I6	0.77	0.01	0.90	0.18	0.03	0.24	2.36
W8	0.50	0.00	0.88	0.20	0.03	0.23	3.25
PU	0.70	0.03	0.33	0.11	0.01	0.05	3.24
SI	0.18	0.00	0.21	0.04	0.01	0.02	4.34
JP	0.54	0.01	0.58	0.13	0.02	0.13	**3.29**
I6	1.60	0.05	1.45	0.72	0.09	0.56	2.70
W8	0.90	0.05	2.06	0.68	0.06	0.33	2.29
PU	1.39	0.02	0.88	0.19	0.02	0.12	3.31
SI	0.71	0.06	0.53	0.20	0.02	0.26	3.22
CN	**1.15**	**0.04**	**1.23**	0.45	0.04	**0.32**	2.88

ⓒ2015 Information Processing Society of Japan

Table 6: Average Number of Rhetoric Devices per Topic

	Average Number of Rhetoric Devices per Topic				# Character
Item	**Metaphor**	**Comparison**	**Sarcasm**	**Rhetoric question**	**Tweet Length**
I6	25.33	33.00	58.00	38.67	122.88
W8	15.33	95.00	24.00	46.33	128.72
PU	15.67	65.00	56.00	37.33	130.48
SI	13.67	3.67	85.67	37.67	129.06
EN	**17.50**	**49.17**	**55.92**	**40.00**	127.78
I6	3.33	31.67	2.00	9.00	56.18
W8	7.67	24.00	3.67	0.00	57.49
PU	7.33	22.33	2.67	2.67	70.78
SI	2.33	7.00	1.33	0.00	85.07
JP	5.17	21.25	2.42	2.92	67.38
I6	7.67	103.00	15.67	39.00	80.00
W8	6.00	60.67	9.67	20.67	69.81
PU	7.33	19.33	12.00	34.33	97.76
SI	3.00	3.67	4.00	19.67	81.05
CN	6.00	46.67	10.33	28.42	82.16

needless to say the message level.

Besides, we find that English users use the most intensifiers and diminisher followed by Chinese users and Japanese users and the number of diminisher are much less the number of intensifier in all the three languages. In addition to the emotional signal situation, it seems that English users tend to release their feeling more thoroughly. In terms of negation, Chinese users used the most followed by English user and then Japanese users. For sub topics, Japanese users speak the most followed by Chinese users and English users. This with the emotional signals number may show that Japanese users focus more on sharing information other than judging.

5.2 Rhetoric Phenomenon

Table 6 showed the average number of rhetoric devices per topic. We found the English users uses rhetorical devices the most and much more than the other two countries, especially in sarcasm. If we look Table 6 horizontally, we can see that Sarcasm ranks first

Table 7: The Global Polarity Distribution of Each Evaluation Object

Topic	# Positive	# Negative	# Neutral	# Undefined	Positive/Negative
English	205	115	101	29	1.78
Japanese	120	218	94	19	0.55
Chinese	188	154	85	25	1.22
I6	171	162	93	24	1.05
English	86	256	91	22	0.34
Japanese	60	249	148	13	0.24
Chinese	83	284	71	17	0.29
W8	76	263	103	17	0.29
English	44	224	142	40	0.20
Japanese	194	60	208	18	3.23
Chinese	184	130	120	16	1.42
PU	141	138	157	25	1.02
English	194	142	88	26	1.37
Japanese	28	33	382	7	0.85
Chinese	92	72	296	10	1.28
SI	105	82	255	14	1.27

ⓒ2015 Information Processing Society of Japan

in the four rhetoric devices in English, Comparison ranks first in both Chinese and Japanese.

Besides the cultural factors, such as language difference and the expressing custom, we think two reasons can contribute to why English contains more rhetorical phenomenon than the other two language. First, the length of tweet. Though all of the tweets are confined to 140 characters, Chinese and Japanese tweet composed by hieroglyphic characters can express more information content than the alphabetic English tweet. This makes it possible for CJ[c] users use twitter as a way of spreading objective information or comprehensive interpretation regarding a certain objective, while apparently 140 alphabetic characters is not well suitable for these. Therefore, in order to fully explain their feelings, users tend to use rhetorical technic. The Second reason relates to the sub topic, namely the materials people are talking about. We find that there was a football game between Scotland and England after the referendum for the Objective 'Scotland Independence'. The fact that Scotland voted no for independence and booing the national alphabetic of British brings them lots of mocking from England fans, and most of them are in sarcastic way. However, this irritant issue didn't get much attention from CJ users because of the distance and the third party stance.

5.3 Global Polarity

Table 7 showed the polarity distribution of each evaluation object. If we look at the P/N value (the degree of love) we found that cultures' sentiment to objects differs significantly. English users enjoy the Iphone6 most, then the Chinese users, while Japanese users aren't keen on iphone6 on the whole. As to Windows 8, all the three cultures get accordance that none of them likes it. For Russian president Putin, the difference came back once again. The Japanese users and English users are the opposite each other. Chinese stands at the pro-center stance. Regarding Scotland independence, although there isn't huge gap among countries in terms of P/N value. We notice that much less users in Japanese and Chinese have an explicit view on the issue than the English users, which means the interest party are more sensitive than the third party.

Moreover, we also found that English text has most amount of undefined numbers followed by Chinese then Japanese. This may mean that the English message are more flexible than the other two or Japanese annotators had the high accordance (we will talk more on this in Section 6).

5.4 Sub-topic Information

Table 8 shows the top ten sub topics for each evaluation object. We found that even though there are difference, products IPhone 6 and Windows 8 share similar sub topics. Users talked on 'phone', 'acquisition', 'Apple', 'camera', 'screen', 'size' regarding Iphone6 and 'user experience', 'updating' and their 'PC', 'laptop', 'tablet' about Windows 8.

On the contrary, sub topics of Putin varies. English users tended to talk general politic issues on 'Russia', 'Ukraine', 'West', 'Europe' and 'world' etc. Japanese users focused more on '柔道', '空手',and think him as '政治家', '大統領',which may explain their affection. Chinese users mentioned the 'APEC' and 'G20' submit meetings more than the other two, because APEC was hold on Beijing and G20 submit was widespread reported. For Scotland Independence the sub topic between third-party countries and interest countries differs largely. The third party countries discuss the issue at a macroscopic level ,such as campaign ,referendum, democracy etc., while the interest party countries focus on two specific things ,such as 'Scotland vs England', 'football' and 'Smith Commission' etc.

Table 8: The Top Ten Sub Topics for Each Evaluation Object

I6						SI					
English		Japanese		Chinese		English		Japanese		Chinese	
phone	41	画面	23	手机	30	England	40	独立投票	103	公投	153
case	27	Apple	20	屏幕	23	Scotland vs England	37	イギリス	59	英国	50
falling & dropping	27	ケース	18	摄像头	17	football	24	イングランド	33	卡梅伦	23
screen	23	iPhone	16	三星	14	game	19	独立運動	23	爱丁堡	20
size	19	アップデート	14	换手机	14	SNP	18	日本	14	英镑	16
acquisition	16	カメラ	14	弯曲问题	13	vote no but booing	17	ウェールズ	13	大英帝国	16
battery	16	携帯	13	手感	13	referendum	16	アイルランド	13	美国	15
Apple	15	アプリ	12	苹果	12	national anthem	14	歴史	12	中国	14
camera	11	写真	11	真机	12	Smith Commission	13	ポンド	11	英格兰	12
bending	8	片手操作	11	乔布斯	11	Scotland fans	12	カタルーニャ	10	民主	11
PU						W8					
English		Japanese		Chinese		English		Japanese		Chinese	
Russia	74	ロシア	64	G20	39	user experience	99	PC	68	用户体验	73
Ukraine	36	日本	36	俄罗斯	34	laptop	41	OS	27	兼容性	36
Russian	19	柔道	31	美国	27	updating	77	タブレット	25	换系统	29
West	15	空手	29	奥巴马	24	PC	38	設定	20	重装系统	24

c CJ means Chinese and Japanese.

ⓒ2015 Information Processing Society of Japan

speech	13	オバマ	19	APEC	23	app	31	アプリ	20	系统性能	19
people	10	シンゾウ	15	个人魅力	20	technical issues	27	デスクトップ	20	系统更新	19
country	9	アメリカ	13	制裁	18	updating issues	25	Win7	18	微软	17
Vladimir	9	キレネンコ	13	中国	16	user dissatisfaction	13	アップデート	15	系统升级	17
Europe	8	政治家	13	乌克兰	15	#Windows10	13	画面	15	软件安装	13
world	8	大統領	12	克里米亚	14	tablet	13	操作性	13	用户习惯	13

6. Agreement Analysis and Deficiencies

In this section, we will discuss the agreement between annotators and how the additional work improves the agreement between annotators.

As described briefly in Section 5, the whole annotation work involves three steps:

- Step 1 Annotator tagging
  This step has been introduced in Section 4.1,18 annotators carry through the tagging work according to the guideline independently. The answer of this round has been analyzed in Section 5, which is first-hand but is likely to be lack of agreement.
- Step 2 Merging and Checking
  The organizer observes the three annotators' answer for the same message collection and discuss with fourth annotator to decide those unclear messages (may have problem) to obtain a merged dataset. This intermediate dataset is called as Pivot Dataset. As we can expect, Pivot Dataset has a better quality than a purely majority decision dataset.
- Step 3 Annotator Modification
  In the Step 3, the same annotator as the Step 1 are asked to do a modification on their answers. The Pivot Dataset was provided as a recommended answer for annotators to compare their answer and see if there is any need to change or not. In this step, we only compare the different part from the Pivot Dataset.

6.1 Kappa Statistic

Table 9 shows the Kappa Statistic of both Step 1 and Step 3.From Table 9, it is very clear that the agreement between annotators are relatively low. Most of them are below 0.6, which implies that the reliability of the first round annotation work is moderate (0.4~0.6) and need to be improved.

However, sentiment annotation is highly dependent on the annotator's understanding of the text. From our experience, by asking annotators to check their answers again doesn't really make progress, because annotators are opt to stay with their old thinking. Unlike open question, comparison can be a more feasible way to check. In this situation, annotators can compare their answers with recommendation answers to quickly and precisely identify their problems.

The first dataset comes to our mind is the machine-merged majority answers, which can be obtained easily. However, we believe it is not sufficient due to two reasons. First, as we have seen from Table 9, the Step 1 agreement between annotators is not high, this will cause the low quality of the majority answer, which will harm the Step 3 modification. Moreover, efficient quality management can't be carried out if the organizer has little touch of the content of the messages. Without a full-scale understanding of the texts, annotators can easily refute organizer by their own inference. Therefore Step 2 is indispensable for both the annotators and the organizer.

Table 9 shows the Kappa Statistics after the step 3.We can see that all the agreement between annotators increased. The lowest Kappa is 0.664(substantially reliable) and the highest goes up to 0.847(almost perfect).Based on this fact, the annotation result is more reliable therefore.

Pivot Data method includes two important ideas. First, it can be introspective .The Pivot Data is made and checked by human being, so unavoidably it will bring errors in. By comparing with the annotators, a wrong answer can be fixed if two or more annotators refuse to change. The other one is this method can be a recursive process. We can go back to Step 2 again, make a new Pivot Data according to the Last Step 3 and have a new Step 3 until it constricts[d] (or a satisfied threshold).

Table 9: Average Kappa Statistic between Three Annotators

	English				Japanese				Chinese			
Topic	Step 1	Step 3	+/-	Size	Step 1	Step 3	+/-	Size	Step 1	Step 3	+/-	Size
I6	0.521	0.762	0.241	449.0	0.600	**0.844**	0.245	433.0	0.444	0.693	0.249	449
W8	0.536	**0.838**	0.302	453.0	0.544	0.777	0.233	463.0	0.509	**0.843**	0.334	455
PU	0.325	0.698	0.374	450.0	0.519	0.748	0.229	453.0	0.617	**0.847**	0.230	444.0
SI	0.456	0.664	0.208	450.0	0.406	0.741	0.334	443.0	0.569	0.799	0.229	469.0
average	0.459	0.741	0.281	450.5	0.517	0.778	0.260	448.0	0.535	0.795	0.261	454.3

d A machine-merged majority decision answers equals the last Pivot Dataset.

6.2 Annotation Difference

By comparing with the Pivot Dataset, obvious errors can be easily revised, such as irreverent messages tagged as emotional ones, mistaking the mood of author as the evolution of the object, incorrect tagging for lacking of background information, mistaking the analysis result as the opinions, misunderstanding of the message, overlooking emotional signals, forgetting tagging the global polarity, etc.

Besides these human errors, there are a couple of kinds of annotation variation that are difficult or unnecessary to unify. In the following situations, the final result should use the majority decision.

- Subjectivity Difference

  Some messages are between subjective and objective, which makes some think it as subjective, while other think it as objective. Example (12) is one of them.

  (12) Bottom line. Until the BBC is brought onboard or booted out Scotland will not gain independence Huge audience believes all that is broadcast

- Relevance Difference

  Some messages could be either relevant or irrelevant. In Example (13), one can think that it's a message on YouTube. On the other hand, one can think it is iPhone6's problem that result in the technical issue.

  (13) Can someone explain to me why YouTube videos can't run fluidly anymore?? Grr, what is this! #iPhone6

- Understanding Difference

  A same word could have different understandings. 'new born baby' in Example(14) can be interpreted as preciousness or fragility.

  (14)When Someone hands you an I phone whiteout a case it feels like your handling a new born baby #iPhone6

- Thinking Difference

  Sometimes to a same message, people could think of it in more than one way. For Example (15), generally one think of it as positive because iphone6 make the author 'cool'. While some people think that be used only as a tool of showing off other than realizing its value as a phone is very sad, which is negative to the phone.

  (15)Think I want to buy an #iPhone6 . not because I like them.. but because apparently it makes me cool.. and I just wanna be cool.. that's all

- Culture Difference

  The background of annotator may influence the tagging task. In Example (16), western people tend to think communist as a negative signal, while an Asian annotator just think it as a general political conception.

  (16) haters am throwin deuces yeah its peace ,coz am chilled lyk a buddhist long live #putin u tha last communist

- Rhetorical Difference

  Unlike other two rhetorical devices, sarcasm and rhetorical question rely most on the sense of annotators. Sometimes, some can smell the irony in the text, others think it is just normal. Example (17) and Example 18 can be understood in both ways.

  (17)I seriously love how huge my phone is. When I talk on it, it takes the whole side of my face. Every time its like getting a hug. #iPhone6

  (18)what if #Putin is doing all this just to make sure no one is stupid enough to want to clean up after him in the next term?

- Weight Difference

  For message that contains positive and negative signals at the same time. People may have different concept on deciding which one is more. For Example (19),some say good aspect is more while others say that the 'pointless' is a conclusion.

  (19)Performance and Safety feel, Of course the proud feel is awesome for #iPhone6. Still phone without charge in it is pointless

- Others

  Besides the above-mentioned cases, there still are some other situations. Here we show two of them.

  ■ Modified Object : worked

  @GabeAul Went through all the steps to fix #Windows10 DRM that worked in #Windows7 and #Windows8, and then some, but no luck. Weird!

  ■ Background Knowledge : linx7

  Trying not to #lol as toms losing it trying to suss his #linx7 #windows8 ☠ #notsomuchofabargainnow

6.3 Deficiencies

Lastly, we discuss the main deficiencies during the annotation work. All the following issues have an undesirable influence on the global polarity judgment, therefore need to be improved.

- Net Slang/Abbreviation Problems

  Net slang and abbreviations are prevalent in social media,

and sometimes it's difficult for annotators to figure out what they mean. The misunderstanding of the content always results in wrong judging. In Example (20), 'is the shit' doesn't mean something is not good. On the contrary it means 'great' in the internet. In Example (21), what 'UA' represents are blurring. These two examples show their bad influence on the polarity. Considering the flexibleness of social media, we allowed annotators to access internet if needed. However, problems still remain owing to working time limitation (annotators are not able to search as much as they want).

(20)You know technology is the shit when someone's granddad be looking to by an #Iphone6 and I ain't talking about Boondocks

(21)@ArianaGicPerry Apparently, #Putin said he left early it was a long flight and he needed more sleep, etc, and and no one is upset over UA!

- Non-Corresponding Patterns
  Although we design many ways (3 emotional tags, 3 degree tags, 1 sub-topic tag and 12 rhetorical polarity tags) to record the emotional patterns entailed in the messages and works for most of the messages, there are still a small portion of messages that are difficult to figure out their tagging, because their expressing patterns have beyond our definition. In Example (22), if we tag 'perversion' as negative, this message will be interpreted as that the author is against Scotland Independence, which is obviously wrong. Literary messages like Example (22) always give us challenge and drive us forward endlessly.

(22) Any version of Scotland whose finances are guaranteed by English banks is a perversion of independence.

- Lack of Predefined Rules
  Annotators may encounter confusing messages while doing. These confusing messages will show up for a couple of times, if they choose different answers, this deteriorates the agreement. For example, 'should we regard purchasing experience as one part of Iphone6?', if we define that purchasing are not included in the evaluation of the object, there will be no problem. If we don't have a predefined rule, for Example (23), how do we deiced the global polarity of the message that simultaneously include issue and solution? However, these rules cannot be found without observing all messages in the dataset, which needs much effort.

(23) Finally got the #iPhone6 talking again... But it still won't let me delete texts w/o jumping thru hoops! But it works again!! :)

7. Conclusion and Future Work

In this paper we introduced the process of construction of a multilingual annotated corpus for deep sentiment understanding in social media. Firstly, we described the data collection and a way of data selection. Secondly, we depicted the annotation tasks in detail and performed a quantity analysis of the annotation result. Based on the result we observed the difference between languages and topics. Thirdly, we discussed the improvement of sentiment annotation by an introspection method. We found that this method can help improve the annotation agreement effectively. To our best knowledge, our work is the first comprehensive rhetoric corpus for distant languages, namely Chinese, English and Japanese, in social media research.
This paper mainly introduced the whole image of construction of corpus. In the future, we will look into the content of the tweet messages based on the Golden Dataset. By analyzing the content of the messages, we can see if there is a common mechanism that can embody the sentiment expression for different languages. Although we supposed in Section 4.2.1 that languages have similar structure at the sentiment level, we will try if we can prove this hypothesis and see how it will influence the system development in our next work.

Acknowledge
This project is supported by fundings from Graduate School of Environment and Information Science, Yokohama National University.

Reference

1) Liu Bing. Sentiment Analysis and Opinion Mining (2012). Morgan & Clay Publishers.
2) Preslav Nakov, Sara Rosenthal, Zornitsa Kozareva, Veselin Stoyanov, Alan Ritter, Theresa Wilson. SemEval-2013 Task 2: Sentiment Analysis in Twitter. In Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013), pages 312–320, Atlanta, Georgia, June 14-15, 2013.
3) C´ıcero Nogueira dos Santos .Think Positive: Towards Twitter Sentiment Analysis from Scratch. Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 647–651, Dublin, Ireland, August 23-24, 2014.
4) Bing Xiang, Liang Zhou .Improving Twitter Sentiment Analysis with Topic-Based Mixture Modeling and Semi-Supervised Training. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 434–439, Baltimore, Maryland, USA, June 23-25 2014.
5) Huifeng Tang, Songbo Tan, Xueqi Cheng. A Survey on Sentiment Detection of Reviews. Expert Systems with Applications. Volume 36, Issue 7, September 2009, pages: 10760–10773.
6) Alexandra Balahur, Marco Turchi. Improving Sentiment Analysis in Twitter Using Multilingual Machine Translated Data.in Proceedings of Recent Advances in Natural Language Processing, pages 49–55,
Hissar, Bulgaria, 7-13 September 2013.
7) Svitlana Volkova, Theresa Wilson, David Yarowsky. Exploring Demographic Language Variations to Improve Multilingual Sentiment Analysis in Social Media .In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1815–1827, Seattle, Washington, USA, 18-21 October 2013.
8) XIE Lixing,ZHOU Ming ,SUN Maosong. Hierarchical Structure Based Hybrid Approach to Sentiment Analysis of Chinese Micro Blog

ⓒ2015 Information Processing Society of Japan

and Its Feature Extraction. Journal of Chinese Information Processing. 2012-01.
9) Michael Wiegand, Alexandra Balahur, Benjamin Roth and Dietrich Klakow, Andres Montoyo. A Survey on the Role of Negation in Sentiment Analysis. Proceedings of the Workshop on Negation and Speculation in Natural Language Processing, pages 60–68, Uppsala, July 2010
10) Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan .Thumbs up? Sentiment Classification using Machine Learning Techniques. Proceedings of EMNLP, pages. 79--86, 2002
11) Janyce Wiebe, Theresa Wilson, and Claire Cardie (2005). Annotating expressions of opinions and emotions in language. Language Resources and Evaluation, Volume 39, Issue 2-3, pages 165-210.
12) Liu Bing,Minqing Hu,Junsheng Cheng.Opinion Observer: Analyzing and Comparing Opinions on the Web. WWW '05 Proceedings of the 14th international conference on World Wide Web.Pages 342-351 ACM New York, NY, USA.2005
13) Murthy Ganapathibhotla, Bing Liu .Mining Opinions in Comparative Sentences. Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pages 241–248 Manchester, August 2008.
14) Sara Rosenthal, Alan Ritter, Preslav Nakov, Veselin Stoyanov.SemEval-2014 Task 9: Sentiment Analysis in Twitter. Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 73–80, Dublin, Ireland, August 23-24, 2014.
15) Yi-jie Tang and Hsin-Hsi Chen .Chinese Irony Corpus Construction and Ironic Structure Analysis. Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 1269–1278, Dublin, Ireland, August 23-29 2014.
16) Roberto González-Ibáñez,Smaranda Muresan,Nina Wacholder.Identifying Sarcasm in Twitter: A Closer Look. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: shortpapers, pages 581–586, Portland, Oregon, June 19-24, 2011.
17) Zornitsa Kozareva. Multilingual Affect Polarity and Valence Prediction in Metaphor-Rich. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 682–691, Sofia, Bulgaria, August 4-9 2013.
18) Alexandra Balahur, Marco Turchi. Improving Sentiment Analysis in Twitter Using Multilingual Machine Translated Data. Proceedings of Recent Advances in Natural Language Processing, pages 49–55, Hissar, Bulgaria, 7-13 September 2013.

ⓒ2015 Information Processing Society of Japan