A Method of Topic-Specific Multilingual Data Collecting and Annotation Preparation for Social Media

Lu Yujie†, Sakamoto Kotaro†, Shibuki Hideyuki†, Mori Tatsunori†

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

†Email :{luyujie, sakamoto, shib, mori} @forest.eis.ynu.ac.jp

1 Introduction

Due to 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. As a result, it gives researchers an unprecedented chance to leverage it for the purpose of scientific study and many useful applications have been put forward so far (Liu, 2012).

The dynamic nature of social media with more dynamic ways of expression has implications for the study of sentiment analysis. What's more, social media makes it possible to offer people a collection of multilingual messages. Based on the multilingual setting, researchers therefore become able to investigate the difference between cultures.

The biggest difference between tweet and general text, such as newswire, is tweets are often expressed in a flexible way and the sentiment contained in them are usually subtler. For example:

> *"last #windows8 update took more time than loading 20 #c64games with #datasette ...what went wrong in 30 years?"*

In this example, we can see symbols such as #sign, @sign which are specific symbols only used in social media. These special symbols show the specificity of social media in the expression form. Besides, the author used a rhetoric question at the end of the message to express his discontent with Window8, which needs one more one step of interpretation. To have a better understanding of the expression way users may use, it's necessary to observe the real tweets and annotate them by human beings. If we can show the clues for the sentiment of a message is the following way, it's easier not only for people to read the message but also for machine learning system to learn better.

> *"last #windows8 update[POSITIVE] took more time than loading 20 #c64 games[COMPARISON POSITIVE] with #datasette[SUBTOPIC] ...what went wrong in 30 years?[RHETORIC QUESTION NEGATIVE]"*

Moreover, if we look into the failure cases of the sentiment classification, we will find those sophisticated tweets often contains rhetoric phenomenon. As to the example tweet, it is difficult to decide the polarity of this tweet only based on the sentiment dictionary and words' surrounding features. At this time it's important to either mine out the comparison between windows8 and C64 games or recognize the rhetoric question at the end. We consider rhetoric phenomenon are important in social media and needed to discuss in depth.

In this paper, we discuss the annotation of rhetoric information contained in the tweet. However, although the annotated dataset constructed in this paper is for the use of multilingual sentiment analysis afterwards, we are not going to involve every step of the annotation work for the moment. In this paper, we will focus on the data collecting and the annotation preparation parts.

2 Related Work

In order to evaluate and improve proposed methodologies, annotation is often applied to 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, such as MPQA corpus (Wiebe 2005).

As to social media, a famous one is the dataset offered by the SemEval 2013 Task 2(also 2014) (Preslav Nakov, 2013). SemEval Task 2 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. Many related researches used Task B as their experiment data (Alexandra Balahur 2013; Bing Xiang 2014) for different purposes. Besides Semval, there are I-SIEVE corpus[1], Spanish TASS corpus (Villena-Rom ́an, 2013) etc. For multilingual using, Svitlanna Volkova (2013) constructed a dataset including English, Spanish and Russian by Amazon Mechanical Turk.

Even though 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, as to distant multi-languages such as Chinese, English and Japanese, there isn't an existed annotated dataset for social media[2]. Therefore.it is meaningful to construct such dataset to

[1] http://www.i-sieve.com/

[2] There are parallel corpora (English, Chinese, and Japanese) for machine

help the research on sentiment analysis in multilingual setting in social media.

The most creative part in our annotation project is we will annotated if there is rhetoric phenomenon in the tweet. In linguistics, there are around twenty classes of rhetoric, while in our computational linguistics setting, we only focus on the 4 high-frequently-used rhetoric, including metaphor, comparison, sarcasm and rhetoric question. These four type of rhetoric are defined in a quite loose way, we admit similar rhetoric that is close to these four type. For example, we regard simile, metonymy, and synecdoche as subtypes of metaphor. As stated in the first chapter, by using the features regarding to these rhetoric, the accuracy of sentiment analysis is expected to be improved.

3 Data Construction

3.1 Data Demand

To unveil the different perspectives on a certain object in different regions, the very first step is looking for common or controversial topics discussed between these regions. In our research, we employed 6 topics, including 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

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

3.2 Data Collecting

In this section, we will discuss the data collecting methods for constructing a multilingual corpus including English, Japanese and Chinese. For English and Japanese, we collect the data from Twitter.com by Twitter RESTful API. As to the Chinese, due to the low quality of the messages in Chinese on Twitter, we decide to use data from Weibo.com, which is a Chinese version Twitter.

There are two general ways to collect data regarding an appointed keyword from Twitter: REST Search API and Twitter Streamline API. Twitter Search API returns the similar results as search.twitter.com does, while Twitter Streamline pushes the newest tweets sent by users immediately. According to the different limits for both methods (Table 2), we decide to use the REST Search API in our research.

Table 2 the limitation of REST Search API and Streaming API

	REST Search API	Streaming API
Access Limit	Request limitation are set for each user (the rate limit is based on user, not app)	Single connection for all users (one app only can have one connection)
Time Limit	Maximum back to past 7 days	The instant message
Parameter Limit	Use parameters same as the search.twitter.com	Follow ,Track, Locations Complete matching ;Or relationship; No support for CJK
Rate Limit	100 items per request at most 180 request / 15 minutes (token) 450 request / 15 minutes (App)	Around 3000 items per minute *The real number depends on the parameter combination.
Parameter Limit	Flexible to change Parameters	400 keyword phrases, 5,000 accounts, and 25 geographic areas.

In the implementation of REST Search API, we offer the following functions to guarantee an efficient collecting effect:(1) Automatically set the waiting time if hits rate limiting;(2) Automatic pagination by max_id inherited from Tweepy library; (3) Automatically running every day by registered the task to the operation system;(4) Logging function for check if any error occurs in the process;(5) Complementing tweets if error occurred by source_id. By this mechanism, we can collect around 600 tweet per minute.

As to Weibo.com, the service provider offers a similar RESTful API as Twitter.com does but without a search interface unfortunately, which makes it impossible to collect data by the convenient API.As a result, we resort to the original way --- downloading search result pages directly[3]. In the implementation, we achieved a couple of functions to make it stable and effective:(1) Simulating log-on process using POST data, (2) Auto-pagination by changing the page parameter in the request URL;(3) Repeat one connection for 3 times if timed-out error occurs;(4) Page status confirmation based on file length;(5) Logging function for checking if any error occurs in the process. By this mechanism, we can collect around 75 Chinese tweets per minute (we avoid requesting too frequently.).

3.3 Data Storage

The period we planned to collect is 6 months (2014.10.19~2015.04.18).Table 3 shows the number of data we collected for the first one month. Mention that the maximum number of search result pages on search.weibo.com we can fetch is limited to 50, so we only fetch those original weibos, which mean there is no retweet in the Weibo corpus.

translation use.

[3] s.weibo.com

Table 3 the number of tweets (2014.10.19~2014.11.18)

	English		Japanese		Chinese(original)	
Topic	Total	Ave.(d)	Total	Ave.(d)	Total	Ave.(d)
I6	447525	14436	91659	2957	28208	909
W8	20954	676	10539	340	1624	52
OB	84442	2724	89828	2898	18328	591
PU	433217	13975	148088	4777	24244	782
SI	21422	691	3559	115	743	23
JW	2829	91	30678	990	155	5

Because the return of Twitter API is in JSON type, there is no much need (or not hard) to process them. However, it's necessary to deal with Weibo data for they are contained in HTML file mixed with HTML tags. We design extraction patterns based on the observation of the HTML structure. We then extract all the possible elements from the raw file by regulation expressions and HTML parser, then transfer them into MySQL database. Table 4 shows the data schema of the table for storing them.

Table 4 the database structure for weibo data

Name	Type	Description
Message id	Bingint	The ID of a message.
User id	Bingint	The ID of the sender of the message.
Nickname	Varchar	The nickname of the user.
Identification	Tinyint	The Identification level the user has.
Membership	Tinyint	The membership level the user has.
Is_Taobao	Tinyint	If the user is an online shop owner.
Text	text	The text part of the message.
image number	Tinyint	The number of image
video number	Tinyint	The number of video
column number	Tinyint	The number of information block
sending time	datetime	The sending time of the message.
sending source	Varchar	The platform sending the message.
retweet number	Tinyint	The number of retweet
comment number	Tinyint	The number of comment
thumbup number	Tinyin	The number of thumb-up

4. Annotation Preparation

4.1 Annotation Schema

In order to constructing a Gold Standard for evaluation of the future system. We planned to start an annotation project and build a multilingual corpus. We further choose 4 topics from the 6 topics we collected as our evaluation objects in the annotation stage. For each topic, three different editors will carry through the annotation on it independently according to a common rule set. For each language ,there will be six annotation (Table 5 shows the distribution of the annotators(A1-A6)).To improve the speed and quality of the annotation work, support tool, guidebook and exercise beforehand will be offered or organized before the real work. In this part, we will cover the discussion about the development of the support tool and the selection of data for annotation. The detail of annotation standards and result analysis will be discussed in another work.

Table 5 the distribution of annotators (for one language)

Topic	A.(1)	A.(2)	A.(3)	A.(4)	A.(5)	A.(6)
I6	☑		☑		☑	
W8		☑		☑		☑
PU	☑		☑		☑	
SI		☑		☑		☑

4.2 Annotation Tool

In order to make it more convenient for annotators to tag information to a word, we developed an annotation support tool. Figure 1 shows the mechanism of the annotation tool. The operation file records all the effective operations of the annotator and is stored in JSON format. The cursor events offered by QCursor (Qt Class) take care of the track of the annotators' mouse action.

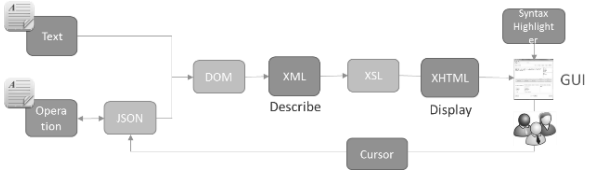

Figure 1 the mechanism of the annotation tool

With the help of the tool, annotators almost don't need to input anything (except for the editing of sub-topics).Basically, all the tasks can be done by mouse and shortcuts. The tool not only support selecting a single word or words in successive order, but also allows annotators to mark up a phrase whose words are far from each other, and automatically generate pair ids (to differentiate the groups in one tweet) with word grouping function. Moreover, we offer view list and information panel to help the annotator know the elements he has tagged in a clear way. The display of a tweet in the text editor changes simultaneously according to user's operation. Figure 2 shows the interface of the annotation tool.

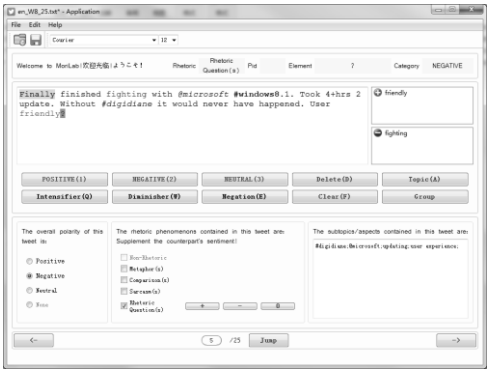

Figure 2 the interface of the annotation tool

4.3 Annotation Dataset

As shown in Table 4, the scale of the corpus is very large, which makes it impossible for people to annotate all 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. These 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 6 shows the examples of the veto patterns used in this stage.

Table 6 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 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 will help us delete a large portion of tweet in the corpus, by which lesson the time and effort for picking up suitable tweets.

5 Discussion

In the preparation stage, we find the following issues can be improved. First, in order to collect data as complete as possible, the query words need to be elaborately designed. An unfit query word may bring you too many useless tweets, which makes it difficult to select data for annotation. In our research we find '#Iphone6' and '#Windows8' in Japanese contains a high percentage of objective tweets. Second, the design of the tool may changes during the development process and the work period. Even though we are able to change its whole interface later, it's necessary to design a compatible data structure for recording operation at first. The change of the data structure will make the new version not compatible with the old operation file.

6 Conclusions

This paper introduced the data collecting and annotation preparation for constructing a multilingual gold standard dataset. The next step for this research is to carry out the annotation work (which is currently undergoing). After finish the work, we will do a detailed analysis on the annotation process.

Acknowledgements

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

References

[1] 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.

[2] 宮崎林太郎, 森辰則. 注釈事例参照を用いた複数注釈者による評判情報コーパスの作成. 自然言語処理, Vol. 17, No. 5, pp. 3−50, (2010).

[3] 筒井貴士, 我満拓弥, 大城卓, 菅原晃平, 永井隆広, 渋木英潔, 木村泰知, 森辰則. 地方議会会議録コーパスの構築および政治情報システム構築を目標としたアノテーションの一提案. 自然言語処理, Vol. 21, No. 2, pp. 125−156, (2014).

[4] 渋木英潔, 中野正寛, 宮崎林太郎, 石下円香, 金子浩一, 永井隆広, 森辰則. 情報信憑性判断支援のための Web 文書向け要約生成タスクにおけるアノテーション. 自然言語処理, Vol. 21, No. 2, pp. 157−212,(2014)

[5] 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.

[6] 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.

[7] 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, pp. 165-210.

[8] Liu Bing. Sentiment Analysis and opinion mining(2012). Morgan & Clay Publishers.