Analysis of Growing Cross-lingual Cascades on Twitter

Hongshan JIN[†]   and   Masashi TOYODA[†]

† The University of Tokyo   4-6-1 Komaba, Meguro-ku, Tokyo, 153-8505 Japan
E-mail: † {jhs, toyoda}@tkl.iis.u-tokyo.ac.jp

Abstract   Recent years, online social network such as Twitter, Facebook has become an indivisible part of our daily life. As a result of easy access and globalization, information propagated quickly and even cross the regions and languages. This paper aims to study the growing cross-lingual information cascades on Twitter. First, we studied 615 million tweets and detected the language distribution of tweets and multilingual users. Then, we measured the size and languages of information cascades and analyzed several features may influence information cascades.

Keyword   Information diffusion, information cascades, cascade growth, cross language

1. Introduction

Social network services have become an important part of our daily life. Take Twitter as an example, by March 2015, there are 302 million monthly active users posting 500 million tweets every day. Also, 77% of the accounts are outside the United States and over 30 languages are supported in Twitter [18]. Similar to Twitter, other popular social medias such as Facebook and Google, have millions of monthly active users and support many kind of languages as well. There is no doubt that these social network services have become more global and multilingual.

With easy access and globalization, Social network services have become new kind of information platforms. On Twitter, contents are shared by users easily and quickly with retweeting and mention functionality and some grow to large information cascades. Accompanied with the cascade growth, there are many hot topics and events propagated across the language and national borders.

"Ice Bucket Challenge", one of the hottest topics in 2014, was an activity involving dumping a bucket of ice water on someone's head to promote awareness of the disease amyotrophic lateral sclerosis (ALS) and encourage donations to research. It went viral on social media during July–August 2014. The hashtag of ice bucket challenge was used all over the world and translated into other languages as well. As a result, this event attracted many participants and increased donations for ALS patients all over the world.

Another example is "Oscars selfie" in 2014, which posted by show host Ellen DeGeneres on her Twitter account. It became the most retweeted message of all time. People reposted and imitated this photo, making it diffused cross regions and languages at amazing speed and size. At the same time, host Ellen DeGeneres's selfie, taken during the broadcast on a Samsung smart phone, resulted to the global marketing effects for Samsung company.

As shown in these examples, accompanied with the growth of contents, information is propagated cross languages and regions. In addition, behind these information diffusion, there exists social influence such as beneficence and commercial. Detecting and analyzing these kind of cross-lingual information diffusion will help to find world news and some social problems. And if we can make use of these cross-lingual information diffusion, it may contribute to beneficence and global marketing. While bunch of research focused on analysis and prediction of these cascade growth, little research is about cross-region and cross-lingual cascades. The goal of this work is to understand and analyze the languages of users, tweets, retweets and mentions. We analyze the factors to influence the cascade grow and cross languages.

The rest of this paper is organized as follows. Section 2 introduces related work and section 3 describes our data crawling and language detection methodology. We conduct basic statistical analysis of information cascades and study the factors behind cascade growth in Section 4. In section 5 we define cross-lingual information cascades and analyze the factors behind them. Finally, Section 6 concludes this work and future work.

2. Related Work

The widespread adoption of online social network services opens a new problem of large-scale information diffusion [11]. Many papers have analyzed and cataloged properties of information cascades, while others have considered predicting the speed, size and structure of cascade growth [11]. Many studies consider the cascade prediction task as a regression problem or a binary classification problem. However, they never considered about the language change during information diffusion.

With the globalization and multilingualism of social network services, recent research has studied language distribution and multilingualism in global social network services [1, 3, 9]. Multiple languages are used in global social network services and on Twitter, only half of the tweets were in English [9]. Hale [1] found 11% of users is multilingual users, who use more than two languages in social network. Social network services although international in scope, is not as multilingual as it might be. It is clear that languages serve as barriers in information diffusion [8]. However, we can observe the cross-lingual information diffusion as well.

Some papers analyzed the role of multilingual users [1, 2] and languages [1, 7, 8] in language communities. The social network analysis of multilingual users indicate us that multilingual individuals could help diminish the segmentation of information spheres online by connecting different language communities [2]. When users do cross languages, [1, 8, 14] suggested these users will engage in larger languages, particularly English. However, previous research did some static analysis of language distribution and language communities, little research is about dynamic language analysis of information diffusion which will be introduced in this paper.

3. Data Collection and Language Detection

3.1. Data Collection

Twitter is one of most global and multilingual social network services and data is publicly available through API. We collected 20.5 million tweets per day, representing 6-7% of all public messages over than 8 years. In the beginning, we collected and broadened the users and tweets from the retweets and mentions of 30 famous Japanese users and their tweets. In this work, we analyzed 615,327,985 tweets and 1,442,263 users from Twitter over one month period in June, 2014. According to previous research, Japanese users and tweets seldom share with people or information in other languages [1, 3]. In our work, we want to testify this assumption as well.

3.2. Language Detection

Tweet Language Detection

We identified the language of each tweet using Language Detection API [16]. Because language identification is difficult on such short text [4], Urls, hashtags, and mentions were temporarily removed from the text of tweets for language detection following the recommendations of Graham, et al. [17]. Also we removed the text containing less than 20 characters and it only cut down 0.8% tweets. We identified 54 languages from the 610 million tweets.

User Language Detection

We detected the languages for each users by statistic of their language usage of tweets. Main language of users were defined by the most frequently used language in their tweets. Table 1 shows the languages of tweets and main language of users, ordered by decreasing number of tweets.

Language	#Tweets	%	#Users	%
English	203662130	33.1	553793	38.4
Japanese	169298415	27.5	413907	28.7
Arabic	47981804	7.8	151929	10.5
Spanish	30706241	4.99	73088	5.07
French	24374187	3.96	63848	4.43
Indonesian	17212893	2.8	55296	3.83
Thai	20970365	3.41	30719	2.13
Portuguese	10889569	1.77	14243	0.99
Korean	8904745	1.45	17691	1.23
Other	75821484	12.3	64573	4.48
Unknown	5506125	0.89	3176	0.22

Table 1 Number of tweets in different languages and number of users with different main languages in Twitter

Owing to the different method to collect the data, the frequency distribution of each language is a little different from previous research [1, 2]. However, the top 10 languages are in line with previous research [1, 2]. In another word, our dataset is quite global and multilingual despite the higher frequency of Japanese.

On Twitter, users can write in many languages and multilingual users are defined as the users who use more than two languages. Given the difficulties with shorter text it is useful to establish a threshold under

which the detection of a language is more likely classifier error than authentic use of the language. For this study, a user was considered as a monolingual user when the proportion of usage of main language in all tweets of this user is at least 80%. Users who use two or more languages in their tweets and usage rate of main language is less than 80%, was classified as a multilingual user. Users with less than four tweets were excluded entirely to avoid having any users in the sample with insufficient data to determine if they are monolingual or multilingual in their Twitter usage. We conducted a human-coding study of a random sample of 100 tweets, and found a substantial agreement between human judges and the language detection algorithm.

Figure 1 shows cumulative distribution function of users' usage rate of main language. The mean value of usage rate of main language is 0.908. Among all users, 17% users meeting our requirement and are considered to multilingual users. When apply the same criteria of multilingual users with Hale's work, the proportion is similar to [1].
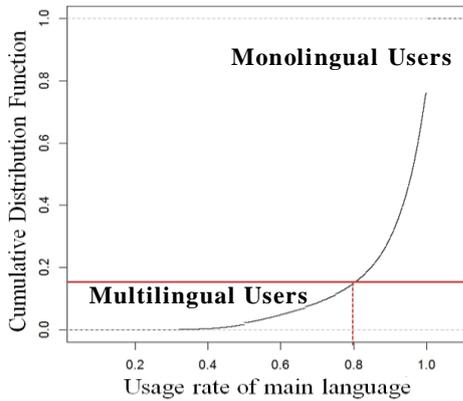

Figure 1 Cumulative distribution function of users' usage rate of main language.

4. Analysis of Cascade Growth

4.1. Information Reshares and Cascades

Twitter allows kinds of convenient conventions. Retweeting is typically used to spread information received from followees to followers [15]. A common form of retweeting is "RT @username message", where "message" is a tweet created by "username". Mentions in the form of @username, allow Twitter users to refer to a specific user. A reply, a specific form of mention with @username appearing at the beginning of the tweet, is a tweet responding to a previous message.

In our study, if user i retweeted or mentioned the tweet of user j, user i is called as resharer and user j is called root user. Similarly, the retweet or mention is called reshare and tweet of a root user is called root tweet. A set of root tweet and reshares is considered as an information cascade and the number of posts in an information cascade is the cascade size. We sampled the information cascades with cascade size. Figure 2 shows that cascade size follows a heavy-tailed distribution. Large scale information cascades are really rare. 94% of cascades are consisted of less than 5 reshares. In other study, cascades with low cascade size are meaningless and we sample cascades with minimum cascade size 10.
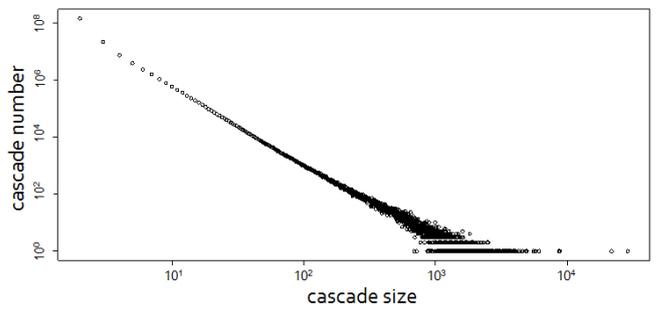

Figure 2 frequency of Cascade with different cascade size

Speed of cascade growth

Information cascades do not grow forever nor stop growing anymore. Found in previous research[11], half of retweeting occurs within an hour, and 75% under a day. However, about 10% of retweets take place a month later. In our work, we observe duration of each reshares and investigate the speed of cascade growth during one month. In figure 3, we can find 94% of reshares occur within 1 day and 98% of the cascades grows within 1 week.
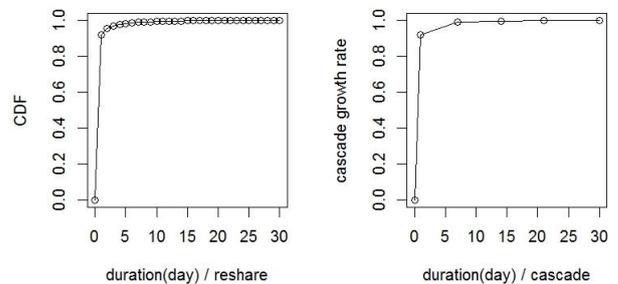

Figure 3 Time duration of reshares and temporal growth of a cascade

Final size of cascade growth

According to previous results, information cascades grows 98% within one week. So in our work, we define one week as information growth duration

and final cascade size within one week as final size f(k). Similar to the previous work[19], we observe the first k reshares of a cascade and observed the final size f(k) with one week. Figure 4 shows the distribution of f(k) with different k reshares. We can find k and f(k) follows linear relations. The median size of final size f(k) is about 1.5 times of k and mean size of f(k) is 2.5 times of k.
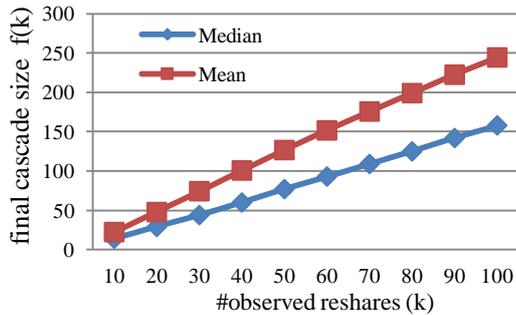

Figure 4 distribution of final cascade size f(k) of observed reshares k

4.2. Factors of Cascade Growth

From the speed and final size observation of cascade growth, we can find most of cascades grow quickly in the first day and tend to stable after one week. Also, half of cascades will reach 1.5 times after first k reshares. Here we try to find what factors will influence the cascades growth.

Previous work attempted on several features of root node and first k nodes, containing content features, root and resharer features, structural features and temporal features [19]. In this study, we testify the basic features of root node and resharers. Here we define activity and influence feature for each user. Activity is defined by the number of prior tweets. Influence is defined by the number of prior reshared tweets. And we define the average number and max number of root user and first k users' activity and influence as cascades' features. Also we add the features of root tweet, containing urls, hashtags, mentions and text length. From correlation coefficient analysis, we find the average activity (0.1113135), average influence (0.106363019) and text length (0.1075459) features have positive relation with cascade growth. More feature selection and analysis will be the future work.

5. Analysis of Cross-lingual Cascades

5.1. Cross-lingual Reshares and Cascades

In an information reshare, if the main language of

the resharer is different to the language of the root user, it is considered as a user-wise cross-lingual reshare. And if the language of the reshare differs from the root tweet, it is defined as a tweet-wise cross-lingual reshare. Otherwise, it is called as a monolingual reshare. Table 2 shows the number of reshares between language pairs.

From the table 2, we can find there is no doubt that the monolingual reshares are more than cross-lingual reshares. However, between different language pairs, the frequency of reshares are different. It shows the different correlation between language pairs. For instance, Spanish has a tighter relationship with English than others. On the country, Arabic has less relations with Asian languages.

Cross-lingual information cascades are the cascades containing tweet-wise cross-lingual reshares or user-wise cross-lingual reshares. In other word, In this information cascade, there is language of at least one reshare differs from the root tweet's language or the main language of the resharer differs from root user's. Monolingual information cascades are defined as the cascades which do not contain any tweet-wise cross-lingual reshares or user-wise cross-lingual reshares. Cross-lingual ratio is the proportion of cross-lingual reshares in all reshares for one cascade. By observing user-wise and tweet-wise cross-lingual cascades, we find the mean size of cross-lingual ratio is only about 8% and most of cascades are monolingual. What kind of cascades will be cross-lingual cascades?

5.2. Factors of Cross-lingual Cascades

According to some previous research, multilingual users and some larger languages can serve as the bridge between language communities. This section studies on the factors such as languages of root users and root tweets of cascade size large than 100 which may result to the cross-lingual and user-wise cross-lingual information diffusion.

In order to find out the correlation between language of root tweets and multilingual cascades, we analyzed the cumulative distribution of cross-lingual cascades for different language of root users and root tweets.

	English	Japanese	Arabic	Spanish	French	Thai	Indonesian	Korean
English	**18185747**	33114	28325	66486	155829	11229	117037	3317
Japanese	20714	**6918259**	489	284	444	733	540	1614
Arabic	23368	136	**3707047**	606	1588	6	2889	10
Spanish	57542	439	763	**2347730**	38886	85	8405	57
French	114400	623	1481	35705	**1800400**	195	12128	45
Thai	4472	366	15	37	119	**1193952**	127	233
Indonesian	82907	1005	1891	5738	8743	358	**714070**	96
Korean	10001	4680	509	81	242	6038	356	**357923**

Table 2 Number of reshares between language pairs

Figure 5 shows the different distribution of English, Japanese, French and Korean root users' cascades. We can find Japanese users' tweets are more likely to be monolingual but Korean users' tweets seem more cross-lingual in cascades with lager size. The cascades of French users' tweets are almost consist of French as well. By manually analyzing the topics of Korean users' tweets, we find the topics of the cross-lingual cascades of Korean root users are closely related to K-pop and propagated in Thai. The reason for this kind of cross-lingual cascades resulted from the popularity of K-pop in Thailand. Here, we can find there is no direct correlation between the size of languages and cross- lingual or monolingual information diffusion. However, the topics of tweets may be the main factors result to the cross-lingual information diffusion.
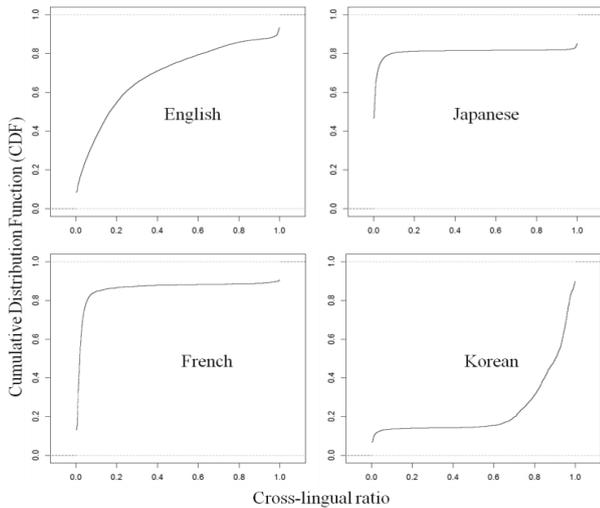

Figure 5 CDF of user-wise cross-lingual ratio for each language of root user.

6. Conclusion

In our work, we studied the language usage on Twitter, especially related to Japanese users. We observed information cascade growth and language change and analyzed the factors behind cascade growth and cross-lingual cascades. Finally, we got the following summarization.

（1）Based on a different dataset which collected from Japanese users, the frequency of languages differed from the previous research. However, the top 10 languages in our dataset is in line with previous work. It indicates our dataset is global and multilingual and suitable for cross-lingual cascade analysis.

（2）On average, about 90% of tweets for a user were posted in a single dominant language. Multilingual users in social networks were less than we expected. It may result from the high thread hold of our definition or the insufficient dataset or just because of their behaviors using the social media.

（3）By analyzing the information reshares and cascades, we find that large cascades are rare and 98% of the reshares grows within one week. In addition, after having observed cascades for one week, half of the cascades grow 1.5 times. We calculated the correlation coefficient between of the factors and cascade size and found users' activity and influence has positive relation with cascade growth.

（4）Most of the information reshares and cascades are monolingual and the mean size of cross-lingual ratio is about 8%. We analyzed the relation between root language, topic and cross-lingual ratio, but the factors to influence the cross-lingual information

diffusion are still unclear.

In our future work, we will aim to analyze more factors behind growing cross-lingual diffusion and set up a prediction model for cross-lingual cascade growth.

Reference

[1] Hale S A. Global connectivity and multilinguals in the Twitter network[C]//Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 2014: 833-842.

[2] Eleta I, Golbeck J. Bridging languages in social networks: How multilingual users of Twitter connect language communities?[J]. Proceedings of the American Society for Information Science and Technology, 2012, 49(1): 1-4.

[3] Hong L, Convertino G, Chi E H. Language Matters In Twitter: A Large Scale Study[C]//ICWSM. 2011.

[4] Papalexakis E, Doğruöz A S. Understanding Multilingual Social Networks in Online Immigrant Communities[C]//Proceedings of the 24th International Conference on World Wide Web Companion. International World Wide Web Conferences Steering Committee, 2015: 865-870.

[5] Papadopoulos S, Kompatsiaris Y, Vakali A, et al. Community detection in social media[J]. Data Mining and Knowledge Discovery, 2012, 24(3): 515-554.

[6] Tang D, Chou T, Drucker N, et al. A tale of two languages: strategic self-disclosure via language selection on facebook[C]//Proceedings of the ACM 2011 conference on Computer supported cooperative work. ACM, 2011: 387-390.

[7] Hale S A. Net Increase? Cross-Lingual Linking in the Blogosphere[J]. Journal of Computer-Mediated Communication, 2012, 17(2): 135-151.

[8] Herring S C, Paolillo J C, Ramos-Vielba I, et al. Language networks on LiveJournal[C]//System Sciences, 2007. HICSS 2007. 40th Annual Hawaii International Conference on. IEEE, 2007: 79-79.

[9] Honey C, Herring S C. Beyond microblogging: Conversation and collaboration via Twitter[C]//System Sciences, 2009. HICSS'09. 42nd Hawaii International Conference on. IEEE, 2009: 1-10.

[10] Marlow C A. The structural determinants of media contagion[D]. Massachusetts Institute of Technology, 2005.

[11] Kwak H, Lee C, Park H, et al. What is Twitter, a social network or a news media?[C]//Proceedings of the 19th international conference on World wide web. ACM, 2010: 591-600.

[12] Takhteyev Y, Gruzd A, Wellman B. Geography of Twitter networks[J]. Social networks, 2012, 34(1): 73-81.

[13] Crystal D. English as a global language[M]. Cambridge University Press, 2012.

[14] Halavais A. National borders on the world wide web[J]. New Media & Society, 2000, 2(1): 7-28.

[15] Hecht B, Gergle D. The tower of Babel meets web 2.0: user-generated content and its applications in a multilingual context[C]//Proceedings of the SIGCHI conference on human factors in computing systems. ACM, 2010: 291-300.

[16] Shuyo. Language Detection Library for Java: https://github.com/shuyo/language-detection

[17] Graham M, Hale S A, Gaffney D. Where in the world are you? Geolocation and language identification in Twitter[J]. The Professional Geographer, 2014, 66(4): 568-578.

[18] Twitter: https://about.twitter.com/company

[19] Yang J, Counts S. Predicting the Speed, Scale, and Range of Information Diffusion in Twitter[J]. ICWSM, 2010, 10: 355-358.

[20] Jenders M, Kasneci G, Naumann F. Analyzing and predicting viral tweets[C]//Proceedings of the 22nd international conference on World Wide Web companion. International World Wide Web Conferences Steering Committee, 2013: 657-664.