一般社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report

Cross-lingual Information Diffusion in Twitter

Hongshan JIN[†] and Masashi TOYODA[†]

[†] Graduate School of Information Science and Technology, 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 With the popularity of social networks service, information propagated quickly across languages and regions. Many researchers studied about the structure of information cascades and predication of cascade sizes. However, despite the global connectivity and multilingualism in these popular social media, there is little research about multilingual social network analysis. This paper aims to study the cross-lingual information diffusion in Twitter, one of the most famous social networks. First, we try to understand and analyze the language distribution of tweets and multilingual users. Then, we analyze the languages of information cascades with large size. Base on these observation and analysis, we propose a cross-lingual model to predict the growth and language distribution of information cascades.

Keywords Information Diffusion, Information Cascade, Multilingualism, Cross languages, Cross region

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 like Facebook and Google, have millions of monthly active users and support many kind of languages as well. There is no doubt that our social networks have become more global and multilingual.

Some previous research proposed that network fragments in communities due to language and national borders [7, 8, 20]. In another word, that social networks always divided into several social communities due to the language and national barriers. In Herring work about language networks on LiveJournal, a kind of social media, he proposed most of the communities consist of one single, dominant language [8].

However, Owing to the global environment and ease of access, there are many hot topics and events propagated across the language and national borders. Ice Bucket Challenge is one of these cross-regional and cross-lingual information. It 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 the 2014 Oscars selfie, 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. Of course it wasn't a marketing stunt. Samsung may have paid a reported $20m for its advertising during the broadcast of the Oscars, but the company, it insisted, was every bit as astonished as everyone else when host Ellen DeGeneres's star-studded selfie, taken during the broadcast on a Samsung smart phone. The amazing speed and size of information propagation resulted to the global marketing effects.

As shown in these examples, social influence like beneficence and commercial effects existed behind these cross-lingual information diffusion. These kind of cross-lingual information diffusion analysis is very necessary and important. How can information propagate across regions and languages? What kind of information will be cross- lingual? The goal of this work is to understand and analyze the languages of users, tweets, retweets and mentions. Then, we analyze the factors to influence the cross-lingual information cascades at a large scale. To the best of our knowledge this work is the first study on the cross-lingual information diffusion in Twitter.

The rest of this paper is organized as follows. Section 2 introduces related work. Section 3 describes our data crawling methodology in Twitter and ’s user profile, trending topics, and tweet messages. We conduct basic statistical analysis of the tweets and users in Section 4. In section 5 we study on factors of multilingual communities and language choice for different topics in social networks. Finally, Section 6 concludes this work and future work.

2. Related Work

With the widespread adoption of social network services, recent research has indicated that multiple languages are used in global social network services. A small qualitative analysis by Honeycutt and Herring [9] found that English, Japanese, and Spanish were the most used languages. Examining 62 million tweets collected over a four-week period, Lichan Hong [3] found that Only half of the tweets were in English (51%). Other popular languages including Japanese, Portuguese, Indonesian, and Spanish together accounted for 39% of the tweets. They first to systematically study the language usage and how users of different languages behave in Twitter. This list is broadly in line with the previous and later studies [1, 9]. It reflects varying uptake of the Twitter platform across the world.

Multilingual users in social networks are defined as the users who use more than two languages and it is useful to establish a threshold under which the detection of a language. Hale [1] studied multilingualism of social networks from the data in Twitter. In their study, user who use two or more languages will be defined as a multilingual user. A user was only considered to use a language if at least 20% of the user's tweets and at least two tweets were detected in that language. All multilingual users, therefore, authored at least four tweets in total. This criteria is similar to the definition given in the research of Irene Eleta [2].

Social network services although international in scope, is not as multilingual as it might be, despite the friendliness of the site design to other languages. It is clear that languages serve as barriers in information diffusion [8]. However, we can observe the cross-lingual information diffusion as well. There is no doubt that someone considered as multilingual users serve as the bridge between language communities [1, 2].

Prevailing theory in social network analysis suggests individuals tend to group together with those similar to themselves. This leads to networks having many clusters or groups of nodes "within which network connections are dense, but between which they are sparser" [1]. These clusters also called as communities result from many factors (gender, race, age, etc.) including language [7, 8]. Here, we aim to find language communities which is a concept in sociolinguistics which means people or a community which uses or shares a common language or maybe a dialect which makes mutual understanding possible or easier among them.

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]. As the discussion in previous section, multilingual users may serve as bridges between communities. [1] illustrated the bridge role of multilingual users as well.

The handful of studies looking at language and social media have found language plays a large role in structuring the hyperlink relationships between blogs [1, 7, 8] and the follower/following relationships between Twitter users [2]. Exactly what information is shared between speakers of different languages on social media and to what extent, however, remains unclear. Multilingual users utilize different languages or mix them for different topics of discussion [4].

When users do cross languages, linguist David Crystal [1, 14] suggests these users will engage with content and users in larger languages, particularly English. Previous studies of language connectivity online have also suggested English plays a special, bridging role connecting speakers of other languages. Herring et al. [8] examined LiveJournal blogs and found language to be a strong factor in structuring 'friend' relationships on the site. English served as a bridging language, and "when non-English journals friend a journal in another language, that language is almost always English." Similarly, Hale [7] examining Japanese, English, and Spanish-language blog posts about the 2010 Haitain earthquake found significantly fewer links between Japanese and Spanish than either Japanese and English or Spanish and English.

As Hale mentioned in his research [1], the factors on cross-lingual information were still unclear. With the different taste of topics, users share information which they are interested in. Similarly, multilingual users may choose some of topics to across languages.

Multilingual users utilize different languages or mix them for different topics of discussion [4, 6]. However, it is not a sufficient work to analyze the relation between topics and multilingualism.

3. Data

3.1. Data Collection

We collected 615,327,985 tweets and 1,442,263 users from Twitter over one month period (June 1 - June 30, 2014). The reason why we choose Twitter is that it is a quite global and multilingual platform and the data is publicly available through API. This data was collected from 2013. In the beginning, we collected and broadened the users and tweets from the retweets and mentions of 30 famous Japanese users and their tweets. On average, we gathered 20.5 million tweets per day, representing 6-7% of all public messages. 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

We identified the language of each tweet using Language Detection API [16]. Because language identification is difficult on such short text [4], Urls, hash-tags, 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. We also detected the languages for each users by statistic 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. 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 multilingual despite the higher frequency of Japanese.

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

4. Analysis

4.1. Multilingual Users

Multilingualism is the use of two or more languages, either by an individual speaker or by a community of speakers. Multilingual speakers outnumber monolingual speakers in the world's population [4]. Owing to the ease of access to kinds of information facilitated by the Internet, multilingualism is becoming increasingly frequent.

Multilingual users in social networks 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].

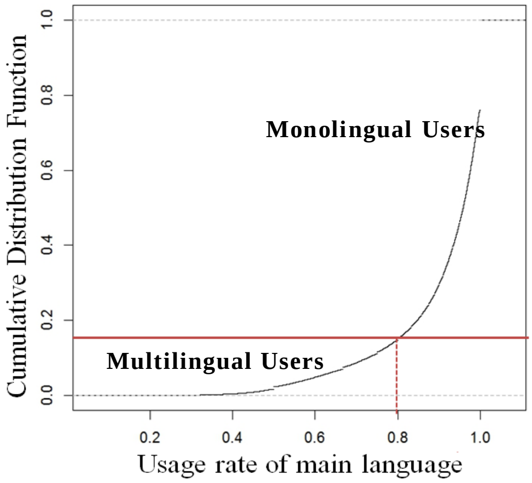

Figure 1 Cumulative distribution function of users' usage rate of main language.

4.2. Multilingual Information Cascades

Twitter allows kinds of convenient conventions. Retweeting is typically used to spread information received from followees to followers [15](Boyd et al. 2010). 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 dataset, there are 32,692,593 retweets/mentions.

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. 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.

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 over 100 and got 59,033 information cascades. 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. The proportion of cross-lingual and monolingual cascades is shown in Table 3.

Cascade types (Cross-lingual ratio)	Tweet-wise	User-wise
Cross-lingual (50%~)	987 (1.7%)	10689 (18.1%)
Cross-lingual (10%~50%)	5194 (8.8%)	7671 (13.0%)
Cross-lingual (~10%)	18483 (31.3%)	22467 (38.1%)
Monolingual (0%)	34369 (58.2%)	18206 (30.8%)
Total	50933	50933

Table 3 frequency of cross-lingual and monolingual cascades with different multilingual ratio.

Less than half of cascades are tweet-wise cross-lingual cascades, however, nearly 70% of the cascades are user-wise cross-lingual cascades. Then what kind of cascades will be multilingual cascades?

	English	Japanese	Arabic	Thai	Spanish	French	Korean	Indonesian
English	**11,820,431**	78,059	12,552	10,710	47,163	81,545	9,304	87,713
Japanese	78,059	**7,565,391**	76	706	5,577	9,581	3,165	32,367
Arabic	12,552	**76**	**3,083,591**	5	1,466	880	229	2,406
Thai	10,710	706	**5**	**1,785,747**	382	500	1,158	1,925
Spanish	47,163	5,577	1,466	382	**1,609,423**	16,994	416	7,649
French	**81,545**	9,581	880	500	16,994	**982,881**	577	5,972
Korean	9,304	3,165	229	1,158	416	577	**740,882**	1,350
Indonesian	**87,713**	32,367	2,406	1,925	7,649	5,972	1,350	**495,093**

Table 2 Number of reshares between language pairs

4.3. Factors on Cross-lingual Information Diffusion

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 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. Figure 2 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. When we analyzed the languages of root tweets, we also got the similar results. 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.

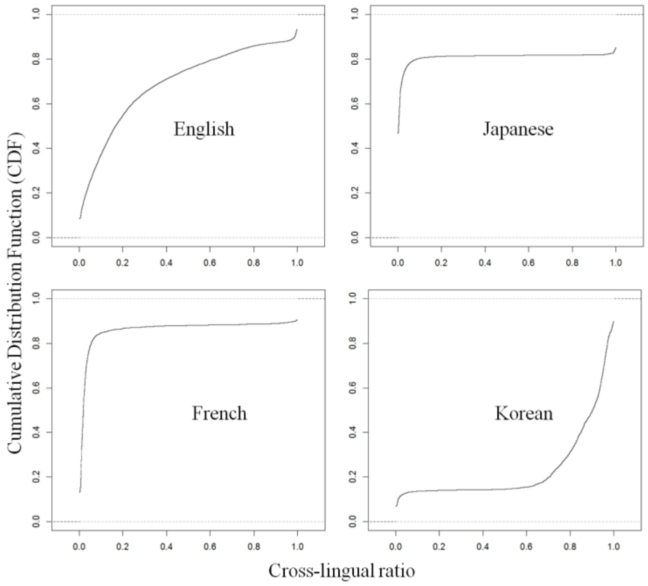

Figure 2 Cumulative distribution function of user-wise cross-lingual ratio for each language of root user.

In order to figure out the correlation between multilingualism of the root users and cross-lingual information diffusion, we calculate the correlation but find that low correlation(0.20) between them. In addition, we find the activities of root users related to the cross-lingual information diffusion. However, other factors like topics of tweets and types of tweets should be analyzed as well.

5. Conclusion

In our work, we studied the language usage in Twitter, especially related to Japanese users. In addition, we analyzed information cascade with large size to figure out the correlation between root users and cross-lingual and user-wise cross-lingual information diffusion. 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.

（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 retweets/mentions network and information cascades, we find that user-wise cross-lingual and cross-lingual information diffusion existed in our social networks. However, the factors to influence the cross-lingual information diffusion are still unclear.

Languages and topics may be important factors when users cross languages. In our future work, we will aim to analyze the factors can cause the cross of languages and regions. Especially, we will focus on the topics and types of information and the information transfer in cross-lingual information diffusion.

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 usage: https://about.twitter.com/company