Detection and Characterization of Influential Cross-lingual Information Diffusion on Social Networks

Hongshan Jin
«Supervised by Masashi Toyoda»
The University of Tokyo
Tokyo, Japan
jhs@tkl.iis.u-tokyo.ac.jp

ABSTRACT

Social network services (SNSs) have become new global and multilingual information platforms due to their popularity. In SNSs with content-sharing functionality, such as "retweet" in Twitter and "share" in Facebook, posts are easily and quickly shared among users, and some of which can spread over different regions and languages. In this work, we first define the concept of cross-lingual information cascade on the basis of the main language of users and then try to characterize and detect those information cascades which can widely spread over different regions and languages on social networks. Understanding the cross-lingual characteristics of information cascades is not only valuable for sociological research, but also beneficial in the practical sense for those who want to know globally-influential events (*e.g.* ALS Ice Bucket Challenge and Terrorism in Europe) and estimate the impact of global advertisements on products (*e.g.* Samsung galaxy phone and a movie, Your Name). On the first attempt, we conducted statistical analysis of cascade growth and language distribution of information cascades with a large Twitter dataset. Based on the results, we propose a feature-based model, by which we successfully detected influential cross-lingual information cascades.

Keywords

Information diffusion; information cascades; cross-lingual cascades; cascade growth; multilingualism

1. PROBLEM

Social network services (SNSs) have become an important part of our daily life due to their widespread adoption. With easy access and less limitation, SNSs have become a new kind of information platform. Posts are shared among users easily and quickly with convenient functions, such as "retweet" and "mention" in Twitter and "share" in Facebook (summarized as reshare). A set of all subsequent reshares starting from the root node that originally create the content is considered as an information cascade (or cascade) [4].

We could find some influential cascades on SNSs. The "ALS Ice Bucket Challenge", one of the hottest topics in 2014, went viral on social media. The hashtag of the ice bucket challenge was used worldwide and translated into other languages. As a result, this event attracted many participants and increased donations for ALS patients worldwide [16]. Another example, the "Oscars selfie" in 2014, became the most retweeted message of all time, which was posted by show host Ellen DeGeneres on her Twitter account. People reposted and imitated this photo, diffusing it across regions and languages at amazing speed and size. At the same time, host Ellen DeGeneres's selfie, taken during the broadcast on a Samsung galaxy phone affected the Samsung's global marketing [15].

In these examples, accompanied with the growth, the cascades were propagated across different languages and regions, which we define as influential cross-lingual information cascades. Understanding the cross-lingual characteristics of these cascades and analyzing the factors behind the information diffusion beyond the language barriers is valuable for sociological research. Based on the analysis, we can predict whether a post can be propagated into any other countries or languages and, if it could, which country or language it can reach to at an early stage. It is significant for many practical applications such as global trend tracking and global marketing. Making use of cross-lingual cascades, we can maximize the social impact and global influence of local events which can promote awareness of social issues (*e.g.* ALS Ice Bucket Challenge) and also improve global marketing (*e.g.* Samsung galaxy phone).

Though there has been a large amount of research on information cascade, much has been focused on just predicting the cascade size, rather than the cross-lingual characteristics. To the best of our knowledge, this is the first work to define and characterize the cross-lingual information cascades, and detect those cascades which will spread widely and be reshared internationally in an early stage. This work is broadly divided into three parts as follows.

- Define cross-lingual/monolingual information cascades on SNSs,
- Observe the cross-lingual characteristics of information cascades to draw out the properties of them,
- Analyze the factors behind influential cross-lingual cascades and build a feature-based model to detect them in an early stage.

WWW'17 Companion, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04.
http://dx.doi.org/10.1145/3041021.3055365

2. RELATED WORK

2.1 Information Cascade

The popularity of online SNSs has resulted in many new research topics of information diffusion [13]. Some researchers analyzed and cataloged properties of information cascades [3][6][13][14], while others considered predicting the speed, size, and structure of cascade growth [1][4].

From the empirical analysis of information cascades on SNSs, some common properties can be observed. Most cascades are small [6][14] and usually occur in a short period of time [3][13].

Based on cascade properties, researchers have attempted to predict the final size of cascades. Many researchers considered the cascade prediction task as a regression problem [1][12] or a binary classification problem [4][12]. One widely used approach to predicting cascade size is the feature-based method [4]. Researchers extracted an exhaustive list of potentially relevant features, mainly including content, original poster, network-structural and temporal features. Then different learning algorithms were applied to predict cascade size. The language distribution of cascades is seldom explored.

2.2 Language Community

With the globalization and multilingualism of SNSs, several recent studies have examined language distribution and multilingualism in global SNSs [5][8]. Social network services are international in scope, but not as multilingual as they should be [7]. Distance and language serve as barriers in social communication [7][9]. They lead to networks having many clusters or groups of individuals with the same language called language communities [9]. Most content is only shared within communities.

Some researchers analyzed the role of multilingual users [5][8][11] and languages [8][9] in language communities. Social network analysis of multilingual users indicates that multilingual individuals could help diminish the segmentation of information spheres by connecting different language communities [5][11]. When users do cross language communities, it was suggested that these users will engage in larger languages, particularly English [8]. These studies inspire us that large languages and multilingual users may contribute to cross-lingual information cascades.

3. APPROACH

In this section, we give the definition of cross-lingual information cascades, then analyze and characterize the cross-lingual information cascades.

3.1 Definition of Cross-lingual Cascade

Since SNSs are global and multilingual, we can access all types of content in different languages. When reshares occur, the users (or resharers) may share the content in a less frequently used language or translate it into their main language (or the most frequently used language). As a matter of fact, reshares with "retweet" or "share" functionality copy the content of the root, thus, not changing the language of the content. We define monolingual and cross-lingual information cascades based on the main language of users which can reflect their languages and regions.

DEFINITION (*Monolingual information cascade*): If the main language of all resharers in a cascade are the same as that of the root user, the cascade is called a monolingual information cascade.

DEFINITION (*Cross-lingual information cascade*): If a cascade contains a resharer whose main language differs from that of the root user, the cascade is considered a cross-lingual information cascade. Accordingly, language distribution of each cascade refers to the fraction of resharers grouped by their main languages in one cascade. The proportion of cross-lingual resharers in a cascade is defined as the cross-lingual ratio r. For a monolingual cascade, r is 0.

3.2 Properties of Cross-lingual Cascade

While there have been many studies to observe the properties including size, speed, and structure of information cascades there has been very less research to analyze the language usage of information cascades. To observe and analyze the language distribution of information cascades on social networks, we choose Twitter as our data source.

Twitter[1] is one of the most global and multilingual SNSs and its data are publicly available through its API. We have crawled more than 2 billion tweets and 1.5 million users from March 1 to July 5, 2014 using the Twitter timeline API. Then we identified the language of each tweet using the Language Detection API[2] developed by Shuyo, which is 99% accurate for 53 languages. We used tweets from March 1 to May 31 to analyze the profile of users and those from June 1 to July 5 to observe the properties of information cascades.

We extracted 74 million information cascades from June 1 to July 5, 2014. The size of information cascades follows a heavy-tailed distribution. The large information cascades are quite rare and only 2% of cascades consisted of more than ten reshares, as proven in previous studies [4][13]. For each cascade, we investigated the speed of cascade growth and found that 98% of cascades grew within one week and tended to stabilize after one week. Then we filtered out about 1 million cascades with the final cascade size $f(k)$ larger than ten and calculated the final cross-lingual ratio $f(r)$ of each cascade during one week. More than half of cascades were monolingual. The mean value of $f(r)$ was only about 11%. It means that predicting cascades with high $f(r)$ is quite difficult.

We investigated the correlation between cascade size and cross-lingual ratio of cascades. We grouped the cascades into the same final cascade size $f(k)$ and calculated the mean value of final cross-lingual ratio $f(r)$. As a result, we found the distribution of cascade size and cross-lingual ratio is independent, and we need to detect influential cross-lingual information cascades by predicting the cascade size and cross-lingual ratio separately.

Similar to the analysis of cascades growth [4], we observed the correlation between the final cross-lingual ratio $f(r)$ and cross-lingual ratio r of the first observed k resharers. Figure 1 shows a box-plot of the $f(r)$ after observing the r of the first ten resharers. We found that the median value of $f(r)$ had a linear relationship (0.9 times) with the r of the first ten resharers. Even if we observe more k, the median value of $f(r)$ shows a linear relationship with the r of the observed k resharers. Only about 20% of cascades would exceed the value of r after observing k resharers. It means that just

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

[2]Library: https://github.com/shuyo/language-detection

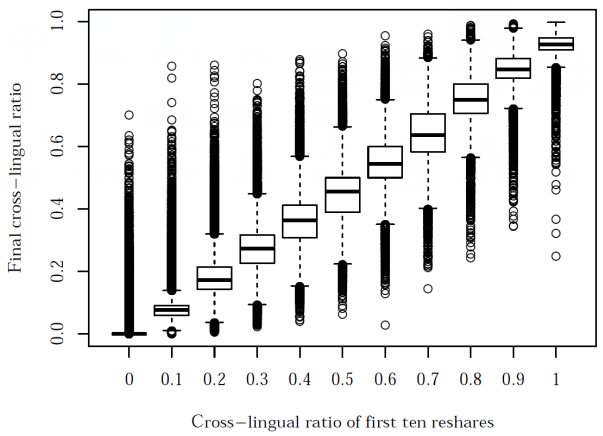

Figure 1: Distribution of final cross-lingual ratio

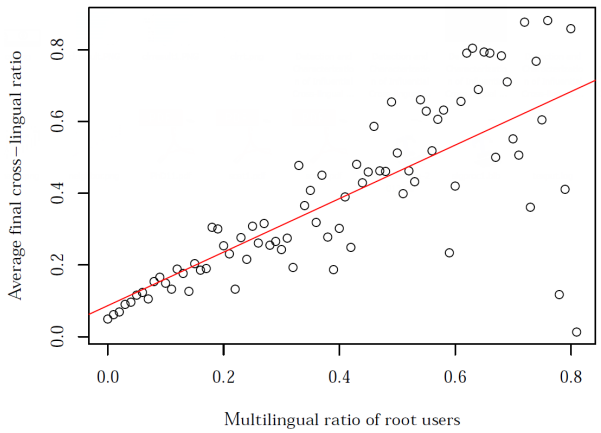

Figure 2: Multilingualism of root users vs. Average of final cross-lingual ratio $f(r)$

keeping r over time is quite difficult. Our target of prediction is whether $f(r)$ exceeds the first observed r.

3.3 Factors behind Cross-lingual Cascade

The feature-based method is one widely used approach to predicting cascade size. As shown in section 3.2, cascade size is not correlated with cross-lingual ratio. To detect cross-lingual cascades, we considered several language related factors of root users and their neighbors.

Effect of Root Users' Main Language

Large languages like English can serve as a bridge between language communities [8]. By connecting language communities, information can spread across language barriers [5]. Accordingly, we assume the root users using different main language have different potential to produce cross-lingual cascades. For different main language users, we investigated the frequency of cascades with a varied range of $f(r)$.

As our assumption, English root users have more cross-lingual cascades than monolingual cascades and cascades with higher $f(r)$ are less. Most cascades from Japanese, Arabic, and Thai speakers are monolingual. Those of root users who speak European languages, Indonesian tend to be more cross-lingual. The main language of root users affects the cross-lingual ratio of their cascades.

Effect of Multilingual Root Users

Since multilingual users may belong to several language communities, they have the potential to propagate information across languages. We investigated the effect of multilingual users on cross-lingual cascades.

Due to the difficulties in language detection for short text, one user is considered to use a particular language when the usage rate of a language is at least 20% and more than four tweets are in that language. A multilingual user uses two or more languages. Among all users in our dataset, 8% met our requirement and were considered multilingual users. The usage rate of languages other than the main language is defined as the multilingual ratio of the user.

We grouped the cascades based on root users' multilingual ratio and calculated the average of $f(r)$. The multilingualism of root users has a positive relation to the $f(r)$ of their cascades, as shown in Figure 2. In other words, cascades from multilingual root users tend to be cross-lingual.

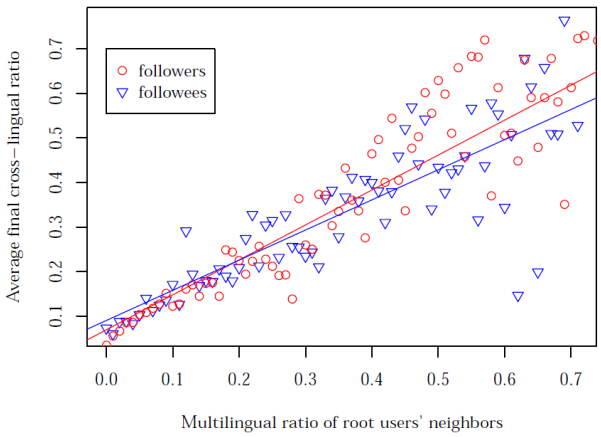

Figure 3: Multilingualism of root users' neighbors vs. Average of final cross-lingual ratio $f(r)$

Effect of Multilingual Neighbors of Root Users

Even though some root users are not multilingual, their tweets can also be cross-lingual and influential if they are internationally famous with followers worldwide. To discuss the influence of international popularity of users on cross-lingual cascades, we analyzed a directed reshare graph extracted from users' previous reshares and determined their monolingual and multilingual neighbors.

Monolingual neighbors refer to neighbors (followers/followees) who share one dominant main language and multilingual neighbors refer to neighbors (followers/followees) who share more than one language and the proportion of the second language is larger than 0.2. The multilingual ratio of neighbors is defined as the proportion of languages other than the dominant main language which reflects the internationality of the user.

We investigated the average $f(r)$ of root users whose neighbors were monolingual and multilingual. Cascades from root users with higher multilingual ratio of neighbors had higher $f(r)$, as shown in Figure 3. In particular, multilingual followers, who represent the international popularity of root users, had higher $f(r)$.

Effect of Content of Root Tweets

The content or the topics of tweets may be considered an important factor affecting cross-lingual cascades. We extracted frequently used words of cascades with different $f(r)$ and in different languages. For instance, for cascades with $f(r)$ larger than 0.8, the main languages were Korean and Thai containing keywords related to famous Korean singers and stars. Cascades with $f(r)$ from 0.2 to 0.7, contained topics related to World Cup 2014 in English and European languages. The top languages used in monolingual cascades were English, Japanese and Arabic. The analysis of root tweets indicates the languages and topics of root tweets are also important for cross-lingual cascade prediction.

4. EVALUATION

Detecting internationally influential information cascades is meaningful and challenging. We first simplified this task as a classification problem to predict cascade size and cross-lingual ratio of cascades. Then we designed and extracted several novel features based on the analysis in Section 3.3, and used machine learning to testify the performance of our feature-based model.

4.1 Problem Formalization

According to previous research [4], we define the cascade size prediction task as a binary classification problem to predict whether the final cascade size $f(k)$ of a cascade reaches the median size during one week after observing the first k reshares of that cascade. For detecting influential cascades, we also consider other classification problems to predict whether the $f(k)$ reaches a specified size such of 100, 500 or 1000.

For predicting cross-lingual cascades, we predicted the final cross-lingual ratio $f(r)$ of cascades. We define the cross-lingual ratio prediction task as a binary classification problem to predict whether $f(r)$ exceeds the r of the first k reshares in one week. As shown in Section 3.2, the fraction of such cascades is only 20%. We evaluated the performance of our prediction model by adjusting the task to predict higher $f(r)$ from lower r.

For the influential cross-lingual cascade prediction task, we considered the multi-classification problem to predict both the size and cross-lingual ratio of cascades. To simplify the evaluation, we define the task as a binary classification problem based on whether $f(k)$ will reach the threshold value and the $f(r)$ will reach r of the first k reshares during one week.

4.2 Features

We now describe the features for prediction. The cascade size prediction problem is not a new topic, and many researchers have proposed several features to predict cascade growth. A previous study [4] showed the importance of structural and temporal features of the root node and first k nodes in a cascade to predict growth. In our influential cross-lingual cascade prediction task, we also used similar features of the root and first k nodes to predict cascade size including root-user, resharer, structural and temporal features. When predicting $f(r)$, language features and content features are important according to the section 3.3. We focus on introducing language and content features in this section.

Language Features

From the previous section, we found that language features of root users are important for cross-lingual cascades. Accordingly, those features of k resharers may also be important. Therefore, we calculated the language features containing the main language, multilingualism, multilingual ratio of the main language, and language distribution vector of the root user and k resharers.

For the root users, we extracted their main language, multilingual ratio and multilingualism of the users and their neighbors (followers/followees). For a more detailed language profile, we include the language distribution vector of tweets and main language distribution vector of their neighbors. For the resharers, we calculated the ratio of multilingual resharers and multilingual neighbors. We also computed the average language distribution vector of their tweets and that of their neighbors.

Content Features

Content is an important feature for cross-lingual cascades, but is less relevant than user features in the cascade size prediction task [1]. We extracted some preliminary features of content, i.e., the language of the root tweet, whether a hashtag, mention, or URL is contained.

To deal with multilingual content data, we trained a topic model based on Wikipedia articles written in the top ten languages used in Twitter. Multilingual articles were grouped into one document by using the inter-language link[3] of articles and modeled by using the Latent Dirichlet Allocation(LDA) topic model [2]. By testing the perplexity of several specified topic numbers, we finally chose 200 as the topic number and inferred the probabilities of topics for each tweet by using this multilingual LDA model.

5. RESULTS

We extracted 1.4 million cascades larger than ten from June 1 to July 5, 2014. As a training set, we randomly sampled 300,000 cascades, the root tweets of which appeared during June 1 to 21. As a test set, we sampled 100,000 cascades, the root tweets of which appeared from June 22 to 28 using simple random sampling algorithm. User and reshare graph features were extracted from March 1 to May 31. We used a linear support vector machine model to conduct the experiments. We trained classifiers on the training set using 10-fold cross validation and evaluated the performance of our model from the accuracy, precision, recall, and F-score on the test set. The baseline classifies all cascade to reach the threshold. The overall performance of our feature-based prediction model for the final cascade size $f(k)$ and the final cross-lingual ratio $f(r)$ prediction tasks after observing ten resharers is shown in Table 1. All the three tasks preforms better than the baseline.

To illustrate the general performance of the features, we contrasted the performance of each feature separately. As shown in Figure 4, language features were significantly better than other features. By correlation coefficient analysis, we found that the multilingual ratio of users' neighbors was the most significant feature. It was followed by the multilingual ratio of the root user and k resharers. Among content features, we found some of the topics, such as music and movies, resulted in cross-lingual information cascades.

[3]https://en.wikipedia.org/wiki/Help:Interlanguage_links

Table 1: Results of influential cross-lingual prediction task after observing ten resharers

$f(k)$	$f(r)$	model	Precision	Recall	F-score
>median	-	baseline	0.51	1	0.67
		our model	0.68	0.78	**0.73**
-	$>r$	baseline	0.17	1	0.29
		our model	0.29	0.71	**0.41**
>median	$>r$	baseline	0.13	1	0.23
		our model	0.27	0.58	**0.37**

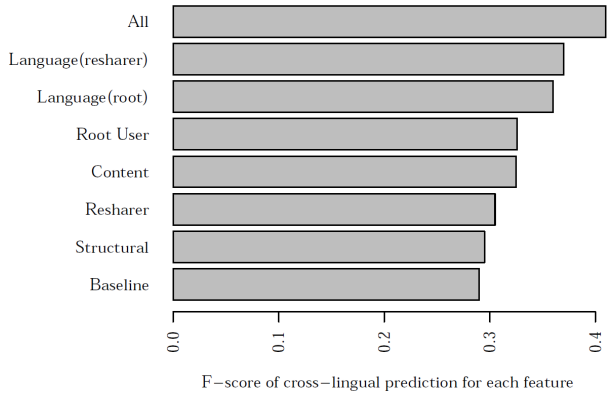

Figure 4: F-score of $f(r)$ prediction for each feature

We examined the sensitivity of prediction performance to the thresholds of cross-lingual ratio. We chose cascades with r less than or equal to 0.1 (the mean value of cross-lingual ratio of cascades), and predicted the performance of our model when changing the threshold value (0.1 and 0.3). Our model performed far better than the baseline, even when the threshold was 0.3. We extensively examined how the prediction performance changed as more resharers observed. Our model showed better prediction performance regardless of k. The performance of the cascade size prediction was slightly improved as k increased.

6. CONCLUSIONS AND FUTURE WORK

We analyzed and detected growing large cross-lingual information cascades on Twitter. It was the first to define the cross-lingual cascade. Cross-lingual cascades, especially with high r were rare and keeping r over time was quite difficult. By analyzing several factors, we found multilingual users and the users with international neighborhood tend to produce cross-lingual cascades. According to these results, we proposed a feature-based model to predict the size and the cross-lingual ratio of information cascades which performed better than the baseline.

This work is at the preliminary stage of detecting influential cross-lingual information diffusion. Deeper observations and better features are required for higher performance of prediction. We will proceed to do deeper analysis in two main areas: improving topic-based language models to deal with multilingual content and extracting structural properties of cascades of differing levels of cross-lingualism.

7. REFERENCES

[1] E. Bakshy, B. Karrer, and L. A. Adamic. Social influence and the diffusion of user-created content. In *Proceedings of the 10th ACM conference on Electronic commerce*, pages 325–334. ACM, 2009.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[3] J. Borge-Holthoefer, R. A. Baños, S. González-Bailón, and Y. Moreno. Cascading behaviour in complex socio-technical networks. *Journal of Complex Networks*, 1(1):3–24, 2013.

[4] J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec. Can cascades be predicted? In *Proceedings of the 23rd international conference on World wide web*, pages 925–936. ACM, 2014.

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

[6] S. Goel, D. J. Watts, and D. G. Goldstein. The structure of online diffusion networks. In *Proceedings of the 13th ACM conference on electronic commerce*, pages 623–638. ACM, 2012.

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

[8] S. A. Hale. Global connectivity and multilinguals in the twitter network. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 833–842. ACM, 2014.

[9] S. C. Herring, J. C. Paolillo, I. Ramos-Vielba, I. Kouper, E. Wright, S. Stoerger, L. A. Scheidt, and B. Clark. Language networks on livejournal. In *System Sciences, 2007. HICSS 2007. 40th Annual Hawaii International Conference on*, pages 79–79. IEEE, 2007.

[10] L. Hong, O. Dan, and B. D. Davison. Predicting popular messages in twitter. In *Proceedings of the 20th international conference companion on World wide web*, pages 57–58. ACM, 2011.

[11] S. Kim, I. Weber, L. Wei, and A. Oh. Sociolinguistic analysis of twitter in multilingual societies. In *Proceedings of the 25th ACM conference on Hypertext and social media*, pages 243–248. ACM, 2014.

[12] A. Kupavskii, A. Umnov, G. Gusev, and P. Serdyukov. Predicting the audience size of a tweet. In *ICWSM*, 2013.

[13] J. Leskovec, M. McGlohon, C. Faloutsos, N. S. Glance, and M. Hurst. Patterns of cascading behavior in large blog graphs. In *SDM*, volume 7, pages 551–556. SIAM, 2007.

[14] J. Leskovec, M. McGlohon, C. Faloutsos, N. S. Glance, and M. Hurst. Patterns of cascading behavior in large blog graphs. In *SDM*, volume 7, pages 551–556. SIAM, 2007.

[15] M. Reed. Who owns ellen's oscar selfie: Deciphering rights of attribution concerning user generated content on social media. *J. Marshall Rev. Intell. Prop. L.*, 14:564, 2014.

[16] L. Townsend. How much has the ice bucket challenge achieved? *BBC News Magazine*, 2014.