Predicting Influential Cross-lingual Information 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 Social network services (SNSs) have become new global and multilingual information platforms due to their popularity. In SNSs with content-sharing functionality, such as "retweets" in Twitter and "share" in Facebook, posts are easily and quickly shared among users, and some grow into large information cascades. Accompanied with such growth, cascades can spread over regions and languages. The "ALS Ice Bucket Challenge" is a good example of internationally influential information that are widely spread and internationally reshared. Analyzing and predicting these influential cross-lingual information cascades will help promote culture exchange and detect international news and issues. This type of research can also provide insights into internal and external relationships among countries and languages. Though there has been a large amount of research on information cascades, much has been focused on predicting just their growth, and little has been on cross-region/lingual information cascades. In this work, we aim to build a robust model and detect influential cross-lingual information diffusion on social networks. To the first attempt, we analyze and predict influential cross-lingual information cascades on Twitter. With a large Twitter dataset, we conducted statistical analysis of growth and language distribution of information cascades. Based on the analysis, we propose a feature-based model to detect influential cross-lingual information cascades and successfully predict their size and language distribution.

Keyword Information cascades; cross-lingual cascades; cascade growth; multilingualism

1. Introduction

Social network services (SNSs) have become an important part of our daily life due to their widespread adoption. Take Twitter as an example. As of June 2016, there were 313 million monthly active users and more than 40 languages supported in Twitter. Other popular social media such as Facebook and Google, have millions of monthly active users and support many languages as well. There is no doubt that these SNSs have become more global and multilingual.

With easy access and less limitation, SNSs have become a new kind of information platform. On Twitter, posts reshared by users easily and quickly with retweeting and mention functionality and some grow into large information cascades. Accompanied with cascade growth, hot topics and events have propagated across languages and national borders. Social network analysts were among the first to show that the social network approach made it possible to follow social ties as they crossed space and language.

The "ALS 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), encouraging donations to research. It went viral on social media from July to August 2014. 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.

Another example is "Oscars selfie" in 2014, which was posted by show host Ellen DeGeneres on her Twitter account. It became the most retweeted message of all time [11]. 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 smart phone affected the Samsung's global marketing.

As shown in those examples, accompanied with growth, information is propagated across languages and regions. Analyzing and detecting these types of influential cross-lingual information diffusion will help detect world issues and social problems. Use cross-lingual information diffusion may contribute to philanthropy and global marketing. While a large amount of research has been focused on analysis and prediction of cascade growth, little has been on cross-region and cross-lingual cascades.

The goal of this study was to detect influential cross-lingual information cascades that will spread widely and be reshared internationally in an early stage. To the best of our knowledge, this is the first study on cross-lingual information diffusion on a large scale and propose a prediction model of cross-lingual cascades.

The rest of this paper is organized as follows. We

introduce related work in Section 2, and describe technical preliminaries based on a large Twitter dataset and define influential cross-lingual information cascades in Section 3. We discuss our basic statistical analysis of information cascades and characterize cross-lingual information cascades in Section 4. In Section 5, we propose our feature-based prediction model to predict influential cross-lingual cascades. Finally, we discuss future work and conclude the paper in Section 6.

2. Related Work

2.1 Information Cascade

The popularity of online SNSs has resulted in the problem of large-scale information diffusion [9]. One of the most widely studied research topics is information cascades. Some researchers analyzed and cataloged properties of information cascades [9][10], while others considered predicting the speed, size, and structure of cascade growth [1][3].

From the empirical analysis of information cascades on SNSs, some common properties can be observed. Most cascades are small [10] and usually occur in a short period of time [9]. Based on cascade properties, researchers have attempted to predict the size of cascades. Many researchers considered the cascade prediction task as a regression problem [1][8] or a binary classification problem [8]. One widely used approach to predicting cascade size was the feature-based method. Researchers extracted an exhaustive list of potentially relevant features, including content, original poster, network-structural, and temporal features [3]. 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 [6][4]. Social network services are international in scope, but not as multilingual as they should be [5]. Distance and language serve as barriers in social communication [5][7]. This leads to networks having many clusters or groups of individuals with the same language called language communities [7]. Most content is only shared within communities.

Some researchers analyzed the role of multilingual users [4][6] and languages [6][7] 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 [4]. When users do cross language communities, it was suggested that these users will engage in larger languages, particularly English [6]. These studies inspire us that large languages and multilingual users may contribute to cross-lingual information diffusion.

3. Definition of Cross-lingual Cascade

Twitter allows convenient functions, such as retweeting and mentioning. Retweeting is typically used to spread information received from followees to followers. Mentions in the form of @username allow Twitter users to refer to a specific user.

Content is shared easily by users using retweets and mentions. The contagious process, in which users reproduce content after having contact with the content influence new users to do the same, is defined as a cascade. We define information reshare and information cascades as follows.

DEFINITION ***Information reshare***: This refers to retweets and mentions. If user u_j retweeted or mentioned a tweet of user u_i, user u_j is called a resharer. Accordingly, the retweet or mention is called the information reshare(or just reshare).

DEFINITION ***Information cascade***: This consists of reshares. A set of all subsequent reshares starting from the root node that originally create the content is considered as an information cascade(or cascade) and the number of reshares in one information cascade is defined as cascade size k. We simply define large cascades as influential cascades.

Since SNSs are global and multilingual platforms, we can access all types of content in different languages. When reshare occurs, resharers may share the content in less frequently used languages or translate them into their native languages. This phenomenon results in information diffusing internationally. 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.

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 number of each main language of resharers in one cascade. The proportion of cross-lingual resharers in a cascade is defined as the cross-lingual ratio.

4. Analysis of Cross-lingual Cascade

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

Twitter 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. Then we identified the language of each tweet using the Language Detection API 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.

4.1 Properties of Cross-lingual Cascade

We extracted 74 million information cascades from June 1 to July 5, 2014. The size of information cascades follow 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 [3][9]. We also investigated how soon a reshare appears and an information cascade grows. By observing the duration of each reshare in June, we found that 94% of reshares occurred within 1 day. For each cascade, we investigated the speed of cascade growth and found that 98% of cascades grew within 1 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 10 and calculated the final cross-lingual ratio *f(r)* of each cascade during one week.

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.

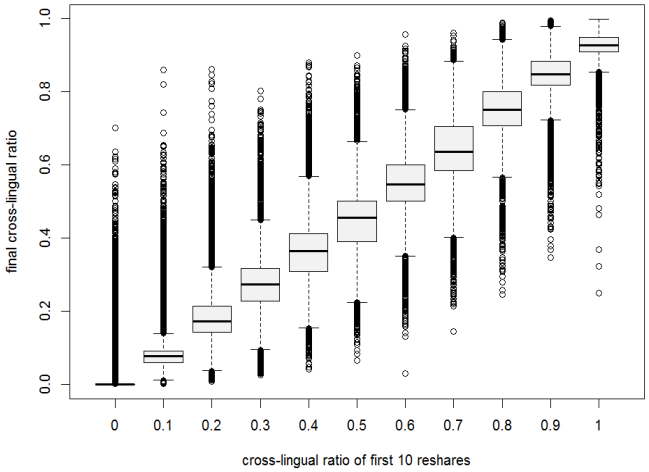

Figure 1 Distribution of final cross-lingual ratio

Similar to the analysis of cascades growth [3], we observed the correlation between the final cross-lingual ratio f(r) and cross-lingual ratio r of the first observed k resharers. We found that the median value of f(r) had a linear relationship (0.9 times) with the r of the first 10 resharers as shown in Figure 1. 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 keeping r over time is quite difficult. Our target of prediction is whether f(r) exceeds the first observed r.

4.2 Factors behind Cross-lingual Cascade

Feature-based method is one of widely used approaches to predict cascade size. In order 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 such as English can serve as bridge among language communities [6]. The main language of root users could be a factor affecting cross-lingual cascades. For different main languages of root users, we investigated the frequency of cascades with different 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. We assume the main language of root users affects the cross-lingual ratio of their cascades.

Effect of Multilingual Root Users

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

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% 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, and cascades from multilingual root users tend to be cross-lingual.

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 reflect 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 a result, multilingual followers, which 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.

5. Prediction of Cross-lingual Cascade

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 extracted several features and used machine learning to testify the performance of our feature-based model.

5.1 Problem Definition

According to previous research [3], 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 specialized 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.

5.2 Feature Extraction

We now describe the features for prediction. Many researchers have proposed several features of the root and the first k nodes to predict cascade growth. Combined with the features proposed in a previous study [3], we designed several novel language related features according to the analysis in the section 4. Finally, we grouped the features into six types: root-user, resharer, content, structural, temporal, and language features.

To deal with multilingual content data, we trained a topic model based on Wikipedia articles. Multilingual arti-

Table 1: Results of influential cross-lingual prediction task after observing 10 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**

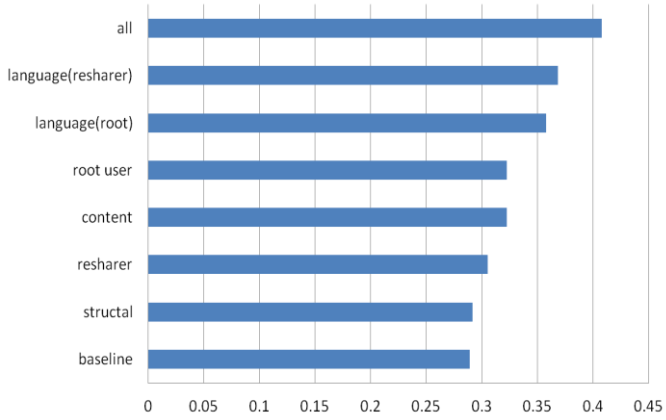

Figure 2: F-score of f(r) prediction for each feature

cles were grouped into one document by using the inter-language link2 of articles and modeled using the Latent Dirichlet Allocation(LDA) topic model [2]. We specified the topic number as 200 and inferred the probabilities of topics for each tweet by using this multilingual LDA model.

Language features contained the main language, multilingualism, multilingual ratio of the main language, and language distribution vector of the root user and k resharers. For k resharers, we calculated the ratio of multilingual resharers and their multilingual neighbors. We also computed the average language distribution vector of their tweets and that of their neighbors.

5.3 Experimental Results

We extracted 1.4 million cascades larger than 10 from June 1 to July 5, 2014. As a training set, we randomly sampled 300,000 cascades, the root tweets of which appear 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. 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 10 resharers is shown in Table 1. All the three tasks performs better than the baseline.

To illustrate the general performance of the features, we contrasted the performance of each feature separately. As shown in Figure 2, 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.

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, and predicted the performance of our mode 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. Conclusion

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. Based on analysis of properties and factors behind cascades, we proposed a feature-based model to predict the size and the cross-lingual ratio of information cascades.

Though the features extracted in our study performed better than the baseline, more detailed importance-performance and error analysis for each feature is required. As mentioned in related studies, feature extraction and extensive training are crucial for this approach, and performance is highly sensitive to the quality of the features. Predicting growing cross-lingual cascades is just the primary stage in detecting influential cross-lingual information diffusion. For future work, we have to consider cascade clustering and information flow between language communities.

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. 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.

[4] 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.

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

[6] 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.

[7] 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.

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

[9] 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.

[10] 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.

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

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