Can Cross-lingual Information Cascades be Predicted on Twitter?

Hongshan Jin[1], Masashi Toyoda[2], and Naoki Yoshinaga[2]

[1] The University of Tokyo, Japan
[2] Institute of Industrial Science, the University of Tokyo, Japan
{jhs, toyoda, ynaga}@tkl.iis.u-tokyo.ac.jp

Abstract. Social network services (SNSs) have provided many opportunities for sharing information and knowledge in various languages due to their international popularity. Understanding the information flow between different countries and languages on SNSs can not only provide better insights into global connectivity and sociolinguistics, but is also beneficial for practical applications such as globally-influential event detection and global marketing. In this study, we characterized and attempted to detect influential cross-lingual information cascades on Twitter. With a large-scale Twitter dataset, we conducted statistical analysis of the growth and language distribution of information cascades. Based on this analysis, we propose a feature-based model to detect influential cross-lingual information cascades and show its effectiveness in predicting the growth and language distribution of cascades in the early stage.

1 Introduction

Online social network services (SNSs) have become more global and multilingual due to their widespread adoption. Let us take Twitter as an example. As of June 2016, there were 313 million monthly active users, 79% of accounts were outsides the United States, and more than 40 languages are supported. Other popular SNSs, such as Facebook and Google+, have millions of monthly active users worldwide and support many languages as well.

With easy access and less limitation, SNSs have become a new type of information platform. Posts are easily and quickly shared among users, with convenient functions, such as "retweet" and "mention" in Twitter and "share" in Facebook (hereafter, referred to as *reshare*). A set of all subsequent reshares starting from the root post that originally created the content is considered an information cascade (or cascade) [4].

Along with cascade growth, some cascades can spread over different regions and languages. We could find several internationally influential cascades. One example is the "ALS Ice Bucket Challenge," which went viral on social media in 2014. The hashtag of the Ice Bucket Challenge was used worldwide and translated into other languages. As a result, this campaign attracted many participants and successfully increased donations for ALS patients worldwide [16]. Another example, the "Oscars selfie," became the most retweeted post in 2014

[15], which was posted by talk show host Ellen DeGeneres on her Twitter account. People reposted and imitated this photo, diffusing them across regions and languages at amazing speed. At the same time, DeGeneres's selfie, taken during the broadcast on a Samsung smart phone affected Samsung's global marketing.

Goal In this paper, we focus on characterizing and detecting influential cross-lingual information cascades that are widely spread and internationally reshared. Though there has been a large amount of research on information cascades, much of the focus has been on just predicting the cascade size and structure [3][4][6][11][12][13]. Language is an interesting and significant, but understudied research topic on information cascades. In this work, we define, analyze and predict growing large cross-lingual information cascades on Twitter. We address the following new problems: what is the linguistic properties of cascades; why and how can information diffuse cross-linguistically and can these information cascades be predicted in an early stage?

Motivation Understanding the cross-lingual characteristics and factors of why information can diffuse beyond language barriers is valuable for sociological research. It can help uncover the global connectivity of social networks and determine relationships among different languages and regions. In addition, accurate prediction of influential cross-lingual information cascades in the early stage is an important enabler of several possible applications. First, knowing the probabilities of information being propagated into difference languages can be used to generate a ranked list of issues for users providing early detection of international breaking news and recommendation of globally-influential events. Second, knowing possible reaction of the multilingual users for a post that advertises a new product, we can quickly evolve (elaborate) the marketing strategy of globally-influential products in the global viral marketing.

Outline of results In this paper, we analyzed the information cascades from 74 million root posts among 1.5 million users based on a six-week data set aggregated from Twitter. We first defined the notion of cross-lingual information cascade on the basis of the main language of root users and resharers, and measured several statistical properties such as cross-lingual ratio (formally defined in § 4.1) of the actual information cascades on Twitter. First, large cascades are rare and 98% of the reshares appear within one week. Accordingly, we focused on the cascades having size larger than ten which grew within one week. After having observed several reshares of cascades, half of them grew 1.6 times irrespective of the observed number of reshares. Second, most of the information reshares and cascades were monolingual. The mean value of the cross-lingual ratio of the cascades was only about 11%. We also found that only a small fraction of cross-lingual information cascades keeps their cross-lingual ratio over time.

Based on the above empirical observations, we define the cascade prediction problem. By analyzing six types of features including content features and language features of the root node and several observed nodes, we propose a feature-based model to predict the size and language distribution of cascades.

Our model detected influential cross-lingual information cascades with better performance than a baseline which uses features for predicting cascade size [4]. The evaluation revealed that language features contributed to the improvement for prediction accuracy both for size and cross-lingual ratio. Specially, multilingual users and users having international followers are likely to produce cross-lingual cascades.

2 Related Work

2.1 Information Cascade

The popularity of online SNSs has resulted in many new research topics of information cascades or diffusion [12]. Some researchers analyzed and cataloged the properties of information cascades [3][6][12][13], while others considered predicting the speed, final size, and structure of cascade growth [1][4].

Existing empirical analysis of information cascades on SNSs [3][6][12][13], revealed common properties of the cascades. Most cascades are small [6][13] and usually grow in a short time [3][12]. Based on these properties, researchers have attempted to predict the final size of cascades. They considered the cascade prediction task as a regression problem [1][11] or binary classification problem [4][11]. One widely used approach to predicting cascade size is the feature-based method [4]. These studies extracted an exhaustive list of potentially relevant features, mainly including content, root user, and network-structural and temporal features. They then applied various learning algorithms to predict cascade size.

Although the increasing multilingual Web indicates the increasing importance of multilingual/cross-lingual studies on information cascades, little work has been done on their cross-lingual behaviors. In this work, we studied cross-lingual characteristics of information cascades.

2.2 Language Community

Several recent studies have examined language distribution and multilingualism in global SNSs [5][8]. Multiple languages are used in global SNSs, and Hale [8] found that 11% of users are multilingual and use more than two languages in Twitter. 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 isolated 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] and languages [8][9] in language communities. Social network analysis of multilingual users indicates that multilingual individuals can help diminish the barriers among different language communities [5]. When users form cross-language communities, these users are likely to engage in larger languages, particularly, English [8]. These studies did not analyze in detail the linguistic influence on information cascades, but have inspired us to argue that large languages and multilingual users may contribute to cross-lingual information cascades.

3 Twitter Datasets

3.1 Data Collection

Twitter is one of the most global and multilingual SNSs and its data is publicly available through Twitter API.[3] We have crawled for more than six years worth of Twitter data using Twitter API from 2011. Our crawling started from 26 famous Japanese users by obtaining their past timelines. We then repeatedly expanded the set of users by following retweets and mentions appearing in their timelines. We continuously expanded users and tracked their timelines. Since we did not limit the language and country of users during expansion, most languages on Twitter were included in the dataset.

Our archive included more than 2 billion tweets and 1.5 million users from March 1 to July 5, 2014. We used tweets from March 1 to May 31 to analyze users and their friendship properties. Based on the user network, we observed the information cascades in tweets from June 1 to July 5.

3.2 Language Detection

Most research on social network analysis has been focused on a single language, regardless of its multilingual characteristic. As a multilingual platform, Twitter has supported more than 40 languages. We identified the language of each tweet using the Language Detection API[4] developed by Nakatani, the precision of which reaches 99% for 53 types of languages on long, clean text such as news articles. Because language identification is difficult in noisy, short text like tweets [14], we temporally removed less language-dependent strings such as URLs, hashtags, and mentions from the text of tweets and detected the languages of tweets containing more than 20 characters. This pre-processing cut only 0.8% of tweets.

After detecting the languages of the tweets, we built a language profile for each user. For each user u_i, we counted the frequency of each language in their tweets and built a language distribution vector, $L_i = (f(l_1), f(l_2), \cdots, f(l_{53}))$. Then we defined the *main language of user* and determined whether a user is multilingual. The main language of a user is defined as the language that is most frequently used in user's tweets. Table 1 shows the language distribution of tweets and languages from March 1 to May 31. The first two columns show the frequency and proportion of the top-10 languages of tweets, ordered by decreasing number, and the distribution of users' main language is shown in the last two columns.

A *monolingual user* is he/she who uses only one language and a *multilingual user* is he/she who uses two or more languages. Due to the difficulties in language detection for short text Twitter, it is useful to set a threshold to determine whether a user uses a certain language. For each user, the *usage rate* of each language is defined as the percentage of that language in the user's tweets. In this study, one user is considered to use a certain language when the usage rate

[3] https://dev.twitter.com/overview/api
[4] https://github.com/shuyo/language-detection

Table 1. Top-10 languages used in tweets and as main languages by users.

Language	# of Tweets(K)	%	# of Users(K)	%
English	534,683	38.8	583	39.2
Japanese	316,902	23.0	433	29.2
Arabic	138,139	10.0	157	10.6
Spanish	81,103	5.89	73	4.92
French	60,263	4.37	67	4.55
Thai	59,101	4.29	32	2.14
Indonesian	41,062	2.98	53	3.55
Portuguese	23,157	1.68	15	0.99
Korean	15,235	1.11	18	1.20
Dutch	15,125	1.10	8	0.56

and number of tweets is more than 20% and 4. The usage rate of languages other than the main language is defined as the *multilingual ratio* of the user. Among all users in our dataset, 8% met our requirement for multilingual users.

Because we initially expand users from 26 Japanese users, the percentage of Japanese tweets was a little higher than that in related studies. However, the top-10 languages were in line with those studies [5][8][10], which shows our dataset is a reasonable subset of Twitter for studying of cross-lingual cascades.

4 Information Cascades

4.1 Definitions

Twitter allows convenient conventions, such as "retweet" and "mention." If user u_j retweeted or mentioned a tweet of user u_i, user u_i is called the *root user* and user u_j is called a *resharer*. Accordingly, the retweet or mention is called the *information reshare* (or *reshare*) and the tweet of the root user is called the *root tweet*. A set of all subsequent reshares starting from the root tweet that originally created 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.

A majority of reshares are done by retweets that just copy the content of the root tweet. Therefore we cannot observe changes in language by retweets even when they are done by foreign users. Instead of observing changes in language, we define the monolingual/cross-lingual information cascades based on the main language of users.

Definition 1. ***Monolingual information cascade*** *If the main languages of all resharers in a cascade is the same as that of the root user, the cascade is called a monolingual information cascade.*

Definition 2. ***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.*

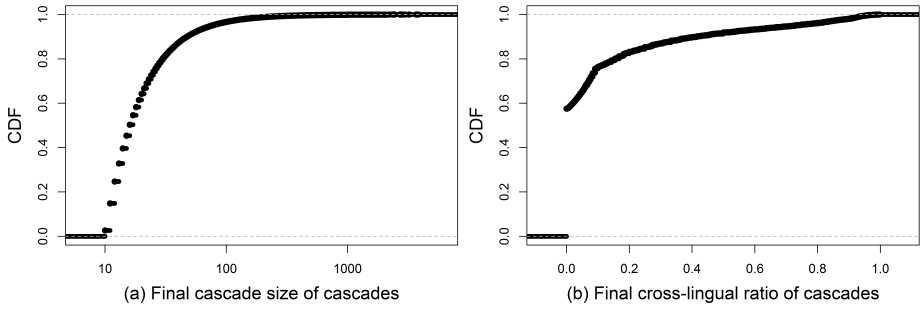

Fig. 1. Cumulative distribution function (CDF) of final cascade size (left) and final cross-lingual ratio of information cascades (right).

Accordingly, the 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* $r_k(>0)$. For a monolingual cascade, r_k is 0.

4.2 Cascade Properties

We extracted 74 million information cascades from June 1 to July 5, 2014 and investigated the distribution of their cascade size. Information cascades follow a heavy-tailed distribution and large information cascades are quite rare, as has been confirmed in the literature [4][12]. 96% of cascades had less than five reshares and only 2% of cascades consisted of more than ten reshares. We investigated how soon information reshares appears and an information cascade grows. By observing the time intervals between each reshare and root tweet in June, we found that only less than 2% of the reshares occurred after one month. For the cascades which root tweets appeared during June 1 to 7, we observed the cascade size during one month and investigated the speed of cascade growth and found that 98% of the cascades grew within one week and tended to stabilize after one week.

Distribution of cascades According to the above size and speed analysis of cascades, we define one week as the duration of cascade growth and define the *final cascade size* $g(k)$ as the size of reshares in a cascade after one week has passed since the root tweet is posted. Similarly, the *final cross-lingual ratio* $f(r_k)$ is ratio of cross-lingual reshares after one week. For the following analysis, we selected about 1.4 million cascades with the final cascade size $g(k)$ larger than ten and the root tweets of which appeared during June 1 to 28 for . The distribution of the $g(k)$ and $f(r_k)$ is shown in Fig. 1.

Fig. 1(a) shows the cumulative distribution function of cascade size. Less than 10% of cascades will exceed 100 and large cascades are really rare. From Fig. 1(b), we can observe that more than half of the cascades were monolingual.

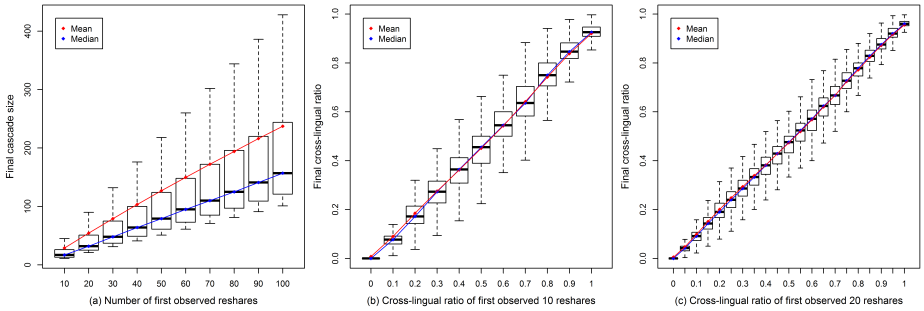

Fig. 2. Box-plot of $g(k)$ and $f(r_k)$ after observing k reshares.

The median value of the cross-lingual ratio was 0 and the mean value of $f(r_k)$ was just 11%. This means that predicting cascades with high $f(r_k)$ is quite difficult.

Final size of cascades In order to detect influential cross-lingual cascades in the early stage, we observed the first k reshares of a cascade and the final cascade size $g(k)$ [4]. This allowed us to study how cascade growth varies with k. From a box-plot of $g(k)$ after observing the first k reshares (Fig 2 (a)), we found that the median value of $g(k)$ is about 1.6 times that of k and the mean value of $g(k)$ is 2.5 times that of k. This indicates that there is a linear relation between the final cascade size and firstly-observed reshares.

Final cross-lingual ratio of cascades Similarly to the analysis of cascade growth, we observed the correlation between the cross-lingual ratio r_k of the firstly-observed k resharers and the final cross-lingual ratio $f(r_k)$. Figure 2 (b) shows a box-plot of $f(r_k)$ after observing the r_k of the firstly-observed ten/twenty resharers. We found that the median value of $f(r_k)$ had a linear relationship (0.9 times) with the r_k of the firstly-observed resharers. Even if we observe more k, the median value of $f(r_k)$ would show a linear relationship with the r_k of the observed k resharers. Only about 20% of cascades will exceed the value of r_k after observing k resharers. It means that maintaining cross-lingual ratio over time is quite difficult. We therefore focus on predicting whether $f(r_k)$ exceeds the cross-lingual ratio r_k based on firstly-observed k reshares..

4.3 Factors Behind Cross-lingual Cascades

§ 4.2 revealed that cross-lingual cascades are rare and predicting large cross-lingual cascades is challenging. In this section, we discuss several factors of root users and tweets that may influence cross-linguality of information cascades.

Influence of root users' main language Commonly used languages such as English can serve as bridges between language communities [8]. By connecting language communities, information can spread across language barriers [5]. We

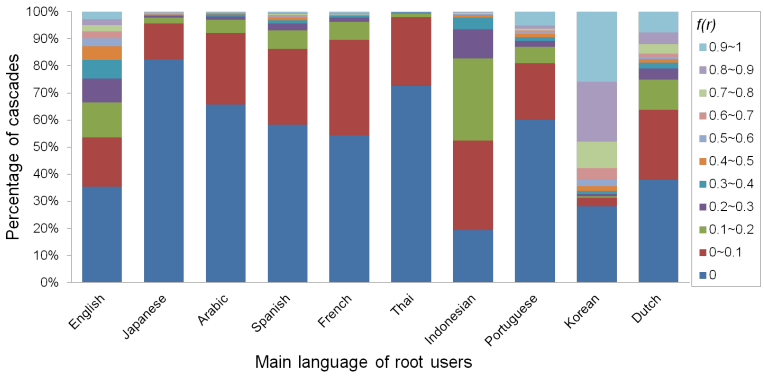

Fig. 3. Main language of root users VS. Cross-lingual ratio of cascades.

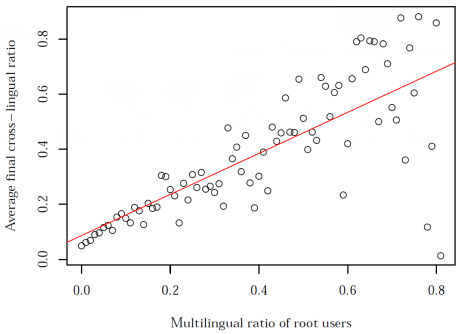

Fig. 4. Multilingual ratio of root users VS. average of $f(r_k)$.

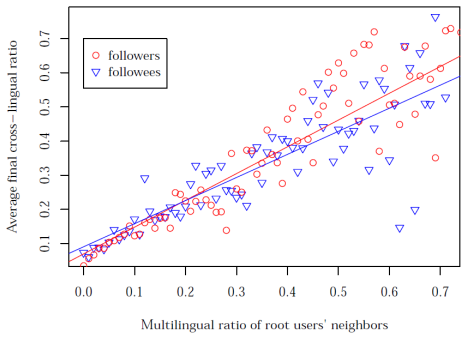

Fig. 5. Multilingual ratio of root users' neighbors VS. average of $f(r_k)$.

assume root users using different main languages have different potential to produce cross-lingual cascades. For each main language of root users, we investigated the frequency of cascades with a varied range of $f(r_k)$ (Fig. 3).

The main language of root users affects the cross-lingual ratio of their cascades. As we had expected, root users whose main language is English produce more cross-lingual cascades than monolingual cascades and cascades with higher $f(r_k)$ are less. Most cascades starting from root users whose main languages are Japanese, Arabic or Thai are monolingual. The cascades of root users whose main language is European languages and Indonesian tend to be more cross-lingual.

Influence of multilingual root users Since multilingual users may belong to several language communities, they have the potential to propagate information across languages. We grouped the cascades by the root users' multilingual ratio and calculated the average $f(r_k)$ of their cascades. The multilingual ratio of root users has a positive correlation to the $f(r_k)$ of their cascades, as shown in Fig. 4. The cascades from multilingual root users tend to be cross-lingual.

Influence of internationally popular root users In some cases, root users are not multilingual, but are internationally famous and have followers worldwide; thus, their tweets can also be cross-lingual and influential. To define the international popularity of users, we analyzed the directed *reshare graph* and defined their *monolingual* and *multilingual neighbors*.

Reshare graph is extracted from users' previous reshares. Nodes represents the root user and resharers. A directed edge e_{ij} represents a user u_j is the resharer of user u_i and u_j is a follower of u_i. We call users connected with each user in the reshare graph as neighbors of the users. Monolingual neighbors refer to neighbors who share one dominant main language and multilingual neighbors refer to neighbors 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 $f(r_k)$ of cascades whose root users' neighbors were monolingual and multilingual (Fig. 5). Cascades from root users with a higher multilingual ratio of neighbors had higher $f(r_k)$. In particular, multilingual followers, which represent the international popularity of root users, had higher $f(r_k)$. By analyzing the effect of the main language of root users and multilingualism of root users and their neighbors, we found that these language factors are important for predicting cross-lingual cascades.

Influence of content of root tweets The content or topics of tweets can be an important factor affecting cross-lingual cascades. We extracted frequently used words of cascades with different $f(r_k)$ and in different languages. For instance, for cascades with $f(r_k)$ larger than 0.8, the main languages were Korean and Thai containing keywords related to famous Korean singers and stars. Cascades with $f(r_k)$ 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 that the languages and topics of root tweets are also important for cross-lingual cascade prediction.

Influence of network structures of root users We investigated the correlation between network structure and cross-lingual ratio of cascades. We extracted an ego network of each root user. Each node in the ego network represents a Twitter user connected to the root user and each weighted, directed edge e_{ij} represents the number of tweets that are authored by user i and mentions and retweets to user j.

We extracted main features including number of nodes/edges, degree assortativity, and density or clusters of the networks, and then calculated the coefficient correlation between these features and $f(r_k)$. As a result, root users who had more denser networks tended to produce cross-lingual cascades.

5 Cascade Prediction

5.1 Problem Definition

Previous studies considered the cascade size prediction task as a binary classification problem [1][11] or regression problem [4][11]. To predict both of the cascade growth and cross-lingual ratio, we define our task as a classification problem and regression problem.

Classification task For the influential cross-lingual cascade prediction task, we define a binary classification problem based on whether $g(k)$ and $f(r_k)$ will reach threshold values after one week. Following the observations in § 4.2, this classification problem is divided into two parts: cascade size prediction and cross-linguality prediction.

Following previous research [4], we define the cascade size prediction task as a binary classification problem to predict whether the $g(k)$ of a cascade reaches the median size (1.6 times of k) after observing the first k reshares of that cascade.

For predicting cross-lingual cascades, we predicted the $f(r_k)$ of cascades. We define the cross-lingual ratio prediction task as a binary classification problem to predict whether $f(r_k)$ exceeds the r_k of the firstly-observed k reshares after one week. As shown in § 4.2, the percentage of such cascades is only 20%. We evaluated the performance of our prediction model by adjusting the task to predict higher $f(r_k)$ from lower r_k.

Regression task Besides predicting $f(r_k)$, we are also interested in predicting the final distribution of languages in cascades. This problem can be solved by a multi-output regression task. The model predicts the final distribution of the resharers' main languages.

5.2 Approach

Our approach to the cascade prediction problem is to represent a cascade by a set of features and then use machine learning methods to predict its future size, cross-lingual ratio and language distribution. For the classification task, we used a linear support vector machine[5] to train the classifier and performed 10-fold cross validation to tune the parameters. In the regression task, we used multi-output regression with random forest and the multi-output meta-estimator.[6] Multi-output meta-estimator performed slightly better than random forest.

We now describe the features for prediction. The cascade size prediction problem is not a new topic, and a previous study [4] showed the importance of structural and temporal features of the root and first k reshares in a cascade to predict growth. When predicting $f(r_k)$, language and content features are important as stated in § 4.3. We grouped the features of the root post and

[5] Liblinear: https://www.csie.ntu.edu.tw/ cjlin/liblinear/
[6] Multi-output regression: http://scikit-learn.org/stable/modules/multiclass.html

firstly-observed k reshares into six types: root-user, resharer, content, structural, temporal, and language (refer to Appendix). We focus on introducing novel language and content features in this section.

Language features From the previous section, we found that the 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. 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 preliminary content features, *e.g.*, the language of the root tweet and whether it contained a hashtag, mention, and URL.

To deal with multilingual content data, we trained a topic model based on Wikipedia articles written in the top-10 languages used in our Twitter datasets. We grouped the multilingual articles into one document by using the inter-language link[7] of articles and modeled using Latent Dirichlet Allocation (LDA) [2]. By testing the perplexity of several specified topic numbers, we finally chose 200 as the number of topics and inferred the probabilities of topics for each tweet by using this multilingual LDA model.

5.3 Experiments

We extracted 1.4 million cascades larger than ten from June 1 to July 5, 2014. As a training set, we sampled more than 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. User and structural features were extracted from March 1 to May 31.

Predicting Influential Cross-lingual Cascades To illustrate the general performance of the features described in the previous section, we observed the root post and first ten reshares of the cascades and predicted whether the final size $g(k)$/final cross-lingual ratio $f(r_k)$ would reach the median $(1.6 * 10)$/cross-lingual ratio r_10 of the firstly-observed ten reshares. We trained classifiers on the training set using 10-fold cross validation and evaluated the performance of our model from the precision, recall, and F_1-score on the test set. The *baseline1*

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

Table 2. Results of influential cross-lingual prediction task after observing 10 resharers.

$g(k)$	$f(r_k)$	Model	Precision	Recall	F_1-score
>median	-	Baseline1	0.42	1	0.60
		Baseline2	0.56	0.78	0.65
		Our model	0.54	0.86	**0.66**
-	$>r_k$	Baseline1	0.17	1	0.29
		Baseline2	0.24	0.63	0.35
		Our model	0.42	0.70	**0.41**
>median	$>r_k$	Baseline1	0.12	1	0.22
		Baseline2	0.27	0.37	0.31
		Our model	0.26	0.59	**0.36**

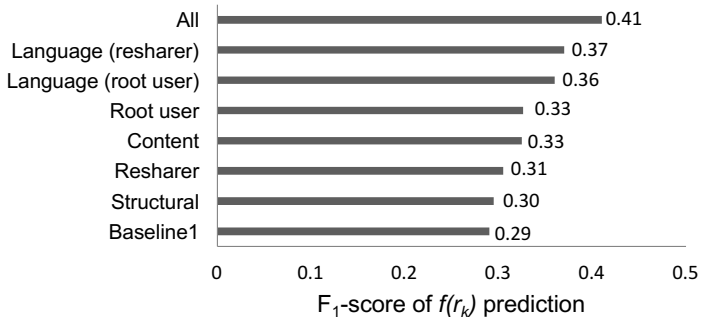

Fig. 6. F_1-score of $f(r_k)$ prediction when using each feature type separately.

(all positive) classifies all cascades to reach the threshold. The *baseline2* uses features for growth prediction and the r_k of the firstly-observed k reshares. The overall performance of our feature-based prediction model for the final cascade size $g(k)$ and the final cross-lingual ratio $f(r_k)$ prediction tasks after observing ten resharers is shown in Table 2. The proposed method performed better than the baselines on all the tasks.

In what follows, we (1) evaluated the importance of each feature, (2) examined the predictability of cross-lingual cascades by changing the threshold of cross-lingual ratio, and (3) examined how the prediction performance of our model changed as more resharers were observed.

Feature importance To evaluate the importance of each feature type, we measured the performance when using each feature type separately. As shown in [4], temporal features most significantly affected the performance of $g(k)$ prediction, followed by structural features. When predicting the $f(r_k)$ of cascades, we found that language features were more effective than other features (Fig. 6). The temporal and structural features were not useful for cross-lingual ratio prediction.

In addition, we measured correlation coefficient between each feature and $g(k)$ or $f(r_k)$. The average time interval between the first half of reshares and average time interval beetween the rest reshares proved to be the most correlated

Table 3. Results of $f(r_k)$ prediction with different threshold.

$f(r_k)$	Model	Precision	Recall	F_1-score
>0.1	Baseline1	0.04	1	0.08
	Baseline2	0.23	0.06	0.09
	Our model	0.29	0.45	**0.35**
>0.3	Baseline1	0.002	1	0.004
	Baseline2	0.22	0.07	0.014
	Our model	0.25	0.21	**0.23**

Table 4. F_1-score of $g(k)$ and $f(r_k)$ prediction with different k.

k	task	Baseline1	Baseline2	Our model
10	$g(k)$>median	0.60	0.65	0.66
	$f(r_k)$ $>r_k$	0.29	0.35	0.42
20	$g(k)$ >median	0.62	0.68	0.70
	$f(r_k)$ $>r_k$	0.37	0.42	0.48
40	$g(k)$ >median	0.61	0.70	0.72
	$f(r_k)$ $>r_k$	0.45	0.48	0.52

with $g(k)$. Among structural features, the total number of unique followers of the root user and first k resharers showed higher relevance than the others. For the $f(r_k)$ of cascades, we found that the multilingual ratio of users' neighbors was the most significant. This was followed by the multilingual ratio of the root user and k resharers. Among content features, we found that some of the topics, such as music and movies, resulted in cross-lingual information cascades.

Sensitivity to threshold values We examined the sensitivity of prediction performance to the thresholds of cross-lingual ratio. In the first experiment, we predicted whether $f(r_k)$ would exceed the observed r_k. In more realistic applications, we would like to know whether $f(r_k)$ would exceed a constant threshold. Since r_k and $f(r_k)$ have a linear relationship, cascades having large r_k values would exceed the threshold with high probability. We thus chose cascades with r_k less than or equal to 0.1, and predicted the performance of our model when changing the threshold value (0.1 and 0.3). As shown in Table 3, our model performed far better than the baseline, even when the threshold was 0.3.

Sensitivity to prediction timing We examined how the prediction performance of our model changed as more resharers were observed. Table 4 lists the F_1-scores when we changed the number of the firstly-observed resharers (k) from 10 to 80. Our model showed better prediction performance regardless of k. The performance of the cascade size prediction slightly improved as k increased, while that of the cross-lingual ratio prediction also improved as k increased, but the improvement from the baselines became smaller. This is because the percentage of cross-lingual cascades tended to increase after observing more resharers, which made prediction easier and decreased the effectiveness of our model.

Table 5. Mean absolute error of estimated language distribution on top-10 languages.

Language	baseline	our model
English	0.072	0.065
Japanese	0.022	0.020
Arabic	0.023	0.019
Spanish	0.019	0.018
French	0.027	0.025
Thai	0.038	0.032
Indonesian	0.014	0.012
Portuguese	0.015	0.013
Korean	0.013	0.012
Dutch	0.006	0.006

Predicting Language Distribution of Cascades In this experiment, a baseline used only the main language distribution of the first ten resharers as features for regression, whereas our model used all the features discussed in § 5.2. We evaluated the performance in terms of the mean absolute error of regression loss for each language. Table 5 shows the language-wise errors of the most frequent languages. Among those top languages, the error of English was the highest, since the most cross-lingual cascades tended to flow into English. Thai showed relatively high error, since most of the cross-lingual cascades that started from Korean tended to flow into Thai. As shown in Table 5, our model could consistently decrease the regression loss of most top languages, and showed the effectiveness of the proposed features.

6 Conclusions

We analyzed and detected influential cross-lingual information cascades using a large dataset on Twitter. We studied the language usage on Twitter and observed the growth and cross-lingual properties of information cascades. We analyzed the factors behind cross-lingual cascades and proposed a feature-based model, which enables the accurate prediction of size and language distribution of information cascades that performed better than the baseline.

This study is just the preliminary stage in detecting influential cross-lingual information diffusion. The prediction performance is still low, which means more error analysis and more detailed observation related to cultural, content, and structural factors is necessary. For future work, we will consider the translated version of content and cluster the cascades written in different languages to paint a more holistic picture of cross-lingual information diffusion.

Acknowledgments This work was supported by the Research and Development on Real World Big Data Integration and Analysis program of RIKEN, and the Ministry of Education, Culture, Sports, Science, and Technology, JAPAN.

A List of features used for learning

Root User Features
whether a user is verified number of friends/followers/followees
number of listed/statues/favorites
number of original/total tweets
number of reshares
number of reshared tweets
Resharer Features
ratio of k resharers who are verified
average/max number of friends of k resharers
average/max number of followers of k resharers
average/max number of listed of k resharers
average/max number of statues of k resharers
average/max number of favorites of k resharers
average/max number of original tweets of k resharers
average/max number of total tweets of k resharers
average/max number of reshares of k resharers
average/max number of reshared tweets of k resharers
Content Features
language of root tweet
whether a hashtag/mention/url is contained
topic distribution of the root tweet
Structural Features
out-degree of root user and kth resharers
in-degree of root user and kth reshares
number of common followers between the root user and kth resharers
total number of unique followers of the root user and k resharers
ratio of k resharers who are not first-degree connections of the root user
Temporal Features
time interval between the root user and kth resharers
time interval between $k-1th$ resharers and kth resharers
average time interval between first half of reshares
average time interval between second half of reshares
Language Features
main language of root user
whether a root user is a multilingual user
usage rate of main language of the root user
whether the follower community of the root user is multilingual
whether the followee community of the root user is multilingual
language distribution of tweets of the root user
main language distribution of followers of the root user
main language distribution of followees of the root user
cross-lingual ratio of k resharers
ratio of k resharers who are multilingual users
ratio of k resharers whose follower community are multilingual users
ratio of k resharers whose followee community are multilingual users
average main language distribution of k resharers' tweets
average main language distribution of k resharers followers
average main language distribution of k resharers followees

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, G. Convertino, and E. H. Chi. Language matters in twitter: A large scale study. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM 2011)*, 2011.
11. A. Kupavskii, A. Umnov, G. Gusev, and P. Serdyukov. Predicting the audience size of a tweet. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*, 2013.
12. H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, pages 591–600. ACM, 2010.
13. J. Leskovec, M. McGlohon, C. Faloutsos, N. S. Glance, and M. Hurst. Patterns of cascading behavior in large blog graphs. In *Proceedings of the 2007 SIAM International Conference on Data Mining*, volume 7, pages 551–556. SIAM, 2007.
14. E. Papalexakis and A. S. Doğruöz. Understanding multilingual social networks in online immigrant communities. In *Proceedings of the 24th International Conference on World Wide Web*, pages 865–870. ACM, 2015.
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.