Predicting Influential Cross-lingual Information Cascades on Twitter
Hongshan Jin* and Masashi Toyoda*
*Univ. of Tokyo  ♠IIS, Univ. of Tokyo

Introduction

- Information can be easily and quickly shared, and some of which can spread over different regions and languages on SNSs
- This is the first study on cross-lingual information cascade on a large scale Twitter data (2 billion tweets and 1.5 million users)
  - Define cross-lingual information cascades
  - Observe the cross-lingual characteristics of cascades
  - Analyze the factors behind influential cross-lingual cascades
  - Build a feature-based model to detect them in an early stage
- Applications: Breaking world news tracking; global marketing

Analysis of Cascades

Definition of Information Cascades:
1. Information cascade: A set of all subsequent reshares (retweets/mentions) starting from the root node that originally create the content
2. Cross-lingual information cascades: A cascade contains a resharer whose main language differs from that of the root user
3. Cascade size (k): number of reshares
4. Cross-lingual ratio (r): proportion of cross-lingual resharsers in a cascade

Properties of Information Cascades:
1. Final cascade size f(k) of cascades (cascades from 6/1 to 7/5, 2014)
2. Final cross-lingual ratio f(r) of cascades
3. Relation between cascade size and cross-lingual ratio

Predicting Cascades

Problem Formulation:
- Influential Cross-lingual Cascade Prediction: a classification task

Proposed approach: feature-based approach
1. Language features
   - Main language of the users
   - Usage rate of main language
   - Multilingual neighbors (followers/followees) of the users
2. Content features: language; topics; length etc.
3. User features: is_verified; #followers; #followees; #friends; #tweets etc.
4. Resharer features: ave(#followers); max(#followers) etc.
5. Structural features: in-degree; out-degree; graph depth etc.
6. Temporal features: time intervals etc.

Evaluation: Linear-SVM
1. Data
   - Train-set: 300,000 cascades (the root tweets appeared 6/1 ~ 6/21)
   - Test-set: 100,000 cascades (the root tweets appeared 6/22 ~ 6/28)
2. Results
   - Influential cross-lingual prediction task after observing 10 resharsers

Future work
1. Improve topic-based language models
2. Extract structural properties of cascades of differing levels of cross-lingualism

Contact jhs@tkl.iis.u-tokyo.ac.jp for more information.