Measuring Discourse Scale of Tweet Sequences: A Case Study of Japanese Twitter Accounts

Shuntaro Yada and Kyo Kaguera
The University of Tokyo
Laboratory of Library and Information Science

UTLiS 東京大学図書館情報学研究室

Introduction

Aim:

To analyse the **discourse scale** of tweet sequences

- A user's tendency of how long she/he tweets successively per coherent discourse

Data:

Tweets of **Japanese** Twitter accounts that declare the followings as their interests:

- Books
- Films

) Similar interests

- Others (interests other than books and films)
  - A representative of average users

Each group has 80 users;

For each user, 3,000 recent tweets and top 50 frequent content words (**discourse keywords**) among them are collected

Motivation:

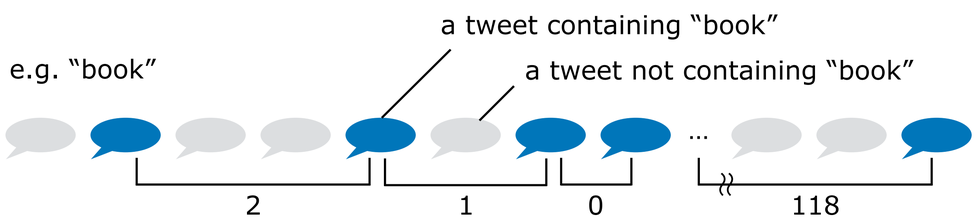
Existing work on Twitter texts has mainly focused on **contents** of tweets (e.g. topic transitions, sentiment analysis, and content analysis). Few studies deal with **formal** aspects of tweets including **discoursal** characteristics. How long users in Twitter talk about their favourite topics in a discourse remains to be clarified.

Also, we are developing a book recommendation system to provide tweets that mention books. The information of the discourse scale can be used for personalisation of the recommendation. For users following those who tend to make long discourses may be likely to accept rich information of books.

Methods and Results

Occurrence Interval

Nearest-neighbour tweet distances of a target word



	Mean	Std
Book	63.17	31.11
Film	64.91	38.60
Others	60.32	27.58

Word	Mean	Std
new title	50.27	149.47
buy	68.14	61.32
library	73.00	138.44
...		

(A part of an actual values of a user; words are translated from Japanese)

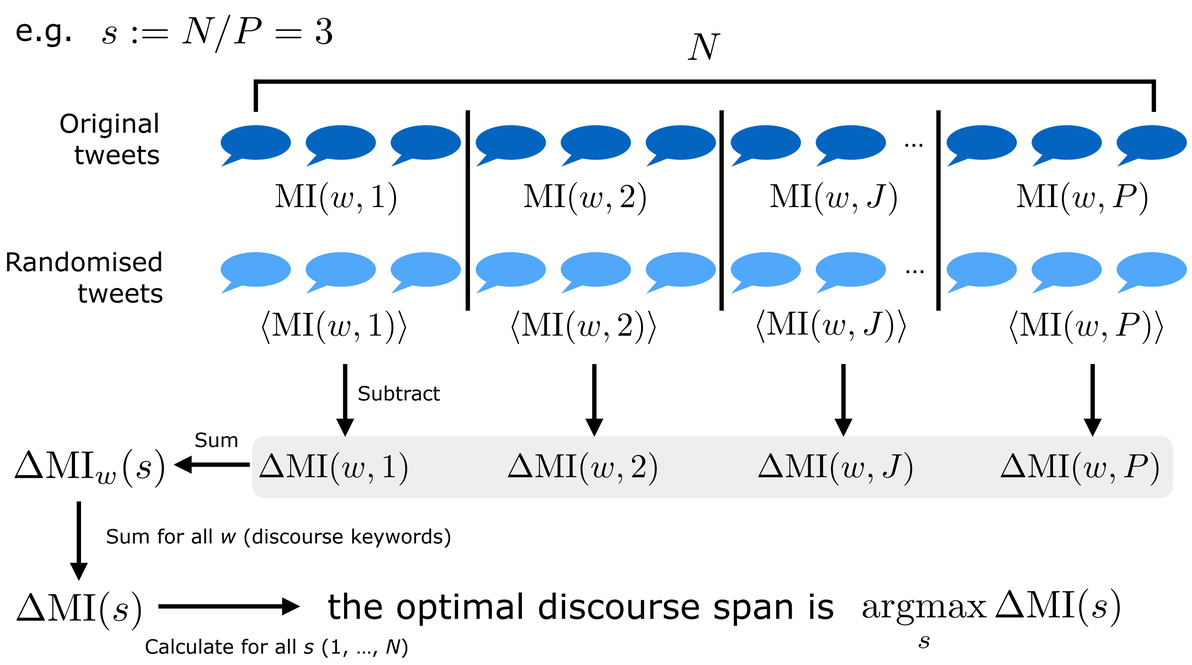
Calculate all successive occurrences of each discourse keyword
e.g. it can reveal a user uses a word every *m* tweets on average

Little outstanding difference between groups

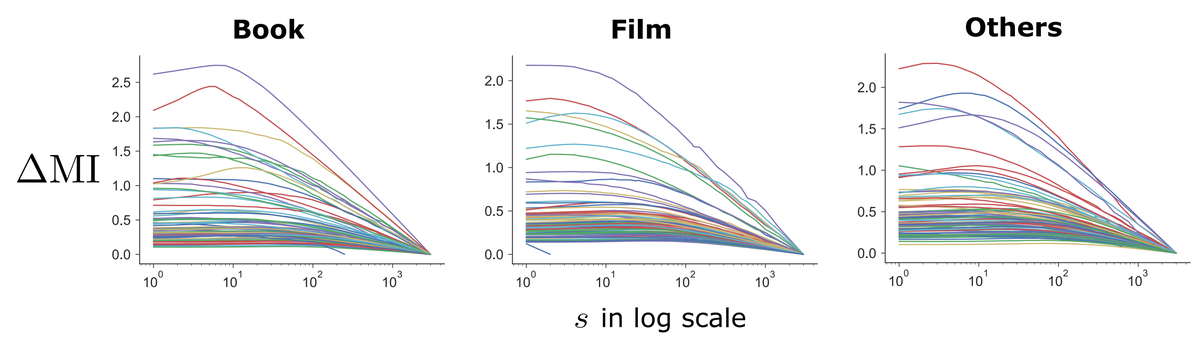
Discourse Span ←Modification of Montemurro & Zanette (2010)

The number of successive tweets that maximise the information amount comparing to random texts of the same length

e.g. $s := N/P = 3$



Discourse span plots per user for each group



Group-wise mean values of optimal discourse spans

	Mean	Std
Book	14.34	12.86
Film	**17.76**	11.28
Others	14.80	14.00

- Large $\Delta\mathrm{MI}$ are produced by bot users
- Film lovers showed larger discourse spans
- Book lovers are similar to average users rather than film lovers

Conclusions

- Film lovers had different characteristics
  - They seem to use more film related words, whereas book lovers mention less book related words
- Discourse spans can be applied for grouping users
- Combining the measure with content-oriented methods will explain users in more detailed way

Feedback Area

This work was supported by JSPS KAKENHI Grant Number JP 16K12542.

ICADL2017, Nov 14th at Bangkok, Thailand