Improving Twitter Retrieval by Exploiting Structural Information

Zhunchen Luo (zhunchenluo@nudt.edu.cn), Miles Osborne (miles@inf.ed.ac.uk)
Sasa Petrovic (sasa.petrovic@ed.ac.uk), Ting Wang (tingwang@nudt.edu.cn)

Introduction

- A tweet can be seen as a structured document constructed from blocks.
  - Plan Text:

    E.g., *Congratulations to the Chinese basketball team for qualifying for the 2012 London Olympics*
  - Text + Link:

    E.g., *Two journalists arrested, in London and Kent, by police investigating alleged corrupt payments to public officials http://bbc.in/NgL5ff*
  - Complex Structures (include hashtag, mention, etc):

    E.g., *This might be the best picture ever. RT @UserA: Us worshipping #GagaTheTalkingBush @UserB pic.twitter.com/zlnofVN8*
- We propose **Twitter Building Blocks (TBBs)** to capture the structural information of tweets to improve ad-hoc Twitter retrieval
- We use a machine learning approach to learn a ranking function for tweets that uses the structural information features (TBB features) and the available social media features.

Twitter Building Blocks

- We define six types of TBBs:
  - TAG: hashtag, e.g., #keywords;
  - MET: mention symbols e.g., @username;
  - RWT: retweet symbols, e.g., RT@username, via @username;
  - URL: links;
  - MSG: content;
  - COM: comment.
- TBB Structures (various combinations of TBBs):

(a) (U need an iphone lol ==>)$_{COM}$ (RT @miiisha_x:)$_{RWT}$ (@XPerkins)$_{MET}$ (i nearly dropped my blackberry in that pooool :()$_{MSG}$

(b) (New iPhone in September – – – – –)$_{MSG}$ (http : / / buswk.co / jbyC0o)$_{URL}$ (# iphone # apple)$_{TAG}$

- Distribution of TBB Structures (2000 gold tagged tweets)

TBB Structures	%	TBB Structures	%
MSG	30.25	TAG MSG	1.55
MET MSG	20.70	TAG MSG URL	1.20
MSG URL	18.40	RWT MSG URL	0.95
OTHERS	13.20	COM RWT MSG	0.85
COM URL	4.10	MET MSG URL	0.85
MSG TAG	2.65	MSG MET MSG	0.70
MSG URL TAG	2.10	RWT MSG TAG	0.70
RWT MSG	1.75		

Automatic TBB Tagger

- Sequential labeling approach (Conditional Random Field).
- Features for TBB tag:
  - Token type;
  - Part-of-speech;
  - Length of the token;
  - Prefix and suffix of characters;
  - Twitter orthography (e.g, The preceding of "RWT" is more likely to be "COM").
- TBB structure identification achieves an accuracy of 82.60%.

TBB Analysis

- It is possible to cluster tweets by TBB structures; these clusters have similar informational characteristics:
  - Public Boardcast: MSG URL; MSG URL TAG and TAG MSG URL
    - E.g., *Apple brings new iPad to China http://bbc.in/NI2HW9*
  - Subjective text: COM RWT MSG (Opinion Retrieval in Twitter. Luo et al, ICWSM-2012)
    - E.g, *I thought we were isolated and no one would want to invest here! RT @UserA: Honda announces 500 new jobs in Swindon.*
  - Messy: OTHERS (the infrequent TBB structures)
    - E.g, *RT @UserA Forreal doeee? (Wanda voic) #Icant cut it out #Newark http://twipic.com/2u15xa ...lmao!!WOW ... http://tmi.me/1UwsA*
- "OTHERS" has the highest Out-of-Vocabulary (OOV) value.

TBB for Learning to Rank Tweets

- TBB features (for example):
  - TBB structure of a tweet (**TBB Structure Type**).
  - The positional information of the query in corresponding TBB.
  - The context information of TBB containing the query.
- Three ranking approaches:
  - Baseline (Duan et al, Coling-2010).
    - Features (for example):
      - **Link: whether the tweet contains a link** (the most effective feature);
      - Length;
      - BM25 score.
  - SM_Rank (More social media features, e.g, number of followers the author of the tweet has).
  - TBB_Rank (Our TBB features).
- Dataset: 100 queries and 936 judged tweets.
- Leaning to rank model: SVM^{Rank}.
- Experimental result (Ten-fold cross-validation):

	MAP		MAP
Baseline	0.4197	Baseline+SM_Rank	0.4546
SM_Rank	0.4338	Baseline+TBB_Rank	0.4326
TBB_Rank	0.4235	SM+TBB_Rank	0.4710
		All	0.4712

- Replace the **Link** feature by **TBB Structure Type** related "URL" block in Baseline.

	MAP		MAP
MSG URL	0.4019	TAG MSG URL	0.3245
MSG URL TAG	0.3327	COM URL	0.3191
RWT MSG URL	0.3289	MET MSG URL	0.1932

Conclusion

- We propose Twitter Building Blocks (TBBs) to capture the structural information of tweets.
- The structural information of tweets can help Twitter retrieval.
- "MSG URL" is the most important structure for Twitter retrieval.