Improving Twitter Retrieval by Exploiting Structural Information

Zhunchen Luo, Miles Osborne, Sasa Petrovic and Ting Wang

Twitter Retrieval

- Most Twitter search systems treat a tweet as a plain text.
- A tweet can be seen as structured text.
- Goal: Improve Twitter retrieval by exploiting structural information.

Structured Tweets

Structured Tweets

Plan Text:

Structured Tweets

Plan Text:

Yao Ming @YaoMing 9月26日

Congratulations to the Chinese basketball team for qualifying for the 2012 London Olympics

Structured Tweets

Plan Text:

Yao Ming @YaoMing 9月26日
Congratulations to the Chinese basketball team for qualifying for the 2012 London Olympics

Text+Link:

Structured Tweets

Plan Text:

Yao Ming @YaoMing 9月26日
Congratulations to the Chinese basketball team for qualifying for the 2012 London Olympics

Text+Link:

BBC News (World) @BBCWorld 53分
Syria says President Assad and UN-Arab League envoy Kofi Annan had a '"constructive and good meeting" bbc.in/MUxICm

Structured Tweets

Plan Text:

Yao Ming @YaoMing 9月26日
Congratulations to the Chinese basketball team for qualifying for the 2012 London Olympics

Text+Link:

BBC News (World) @BBCWorld 53分
Syria says President Assad and UN-Arab League envoy Kofi Annan had a '"constructive and good meeting" bbc.in/MUxICm

Complex Structures (include hashtag, mention, etc):

Structured Tweets

Plan Text:

Yao Ming @YaoMing 9月26日
Congratulations to the Chinese basketball team for qualifying for the 2012 London Olympics

Text+Link:

BBC News (World) @BBCWorld 53分
Syria says President Assad and UN-Arab League envoy Kofi Annan had a '"constructive and good meeting" bbc.in/MUxICm

Complex Structures (include hashtag, mention, etc):

Marc | Lady Gaga @MarcMonster 7月4日
No RT @miiranndaP: New **Lady Gaga**'s dancer. His name is Bernardo Velasco, he's mexican! #PawsUp #HotLikeMexicoRejoice pic.twitter.com/0Qxo3Gsm

Our Work

Our Work

- We propose **Twitter Building Blocks (TBBs)** to capture the structural information of tweets.

Our Work

- We propose **Twitter Building Blocks (TBBs)** to capture the structural information of tweets.

- Learning-to-rank for Twitter retrieval
  - Structural information features (**TBB features**).
  - Social media features (e.g, author social network information).

Twitter Building Blocks (TBBs)

Twitter Building Blocks (TBBs)

- TBB is a sequence of tokens.

Twitter Building Blocks (TBBs)

- TBB is a sequence of tokens.
- Six types of TBBs:

Twitter Building Blocks (TBBs)

- TBB is a sequence of tokens.
- Six types of TBBs:
  - TAG: hashtag, e.g., #keywords.

Twitter Building Blocks (TBBs)

- TBB is a sequence of tokens.
- Six types of TBBs:
  - TAG: hashtag, e.g., #keywords.
  - MET: mention symbols e.g., @username.

Twitter Building Blocks (TBBs)

- TBB is a sequence of tokens.
- Six types of TBBs:
  - TAG: hashtag, e.g., #keywords.
  - MET: mention symbols e.g., @username.
  - RWT: retweet symbols, e.g., RT @username, RT, via @username.

Twitter Building Blocks (TBBs)

- TBB is a sequence of tokens.
- Six types of TBBs:
  - TAG: hashtag, e.g., #keywords.
  - MET: mention symbols e.g., @username.
  - RWT: retweet symbols, e.g., RT @username, RT, via @username.
  - URL: links.

Twitter Building Blocks (TBBs)

- TBB is a sequence of tokens.
- Six types of TBBs:
  - TAG: hashtag, e.g., #keywords.
  - MET: mention symbols e.g., @username.
  - RWT: retweet symbols, e.g., RT @username, RT, via @username.
  - URL: links.
  - COM: comment.

Twitter Building Blocks (TBBs)

- TBB is a sequence of tokens.
- Six types of TBBs:
  - TAG: hashtag, e.g., #keywords.
  - MET: mention symbols e.g., @username.
  - RWT: retweet symbols, e.g., RT @username, RT, via @username.
  - URL: links.
  - COM: comment.
  - MSG: content.

TBB Structures

- TBB structure is a combination of TBBs

TBB Structures

- TBB structure is a combination of TBBs

U need an iphone lol ==> RT @UserB: @UserA

i nearly dropped my blackberry in that poool :(

TBB Structures

- TBB structure is a combination of TBBs

U need an iphone lol ==> RT @UserB: @UserA

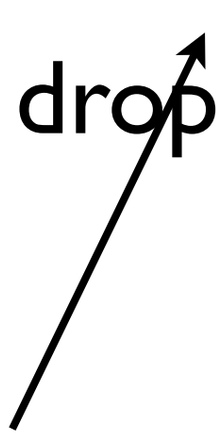
i nearly dropped my blackberry in that poool :(

TBB Structures

- TBB structure is a combination of TBBs

U need an iphone lol ==> RT @UserB: @UserA

i nearly dropped my blackberry in that poool :(

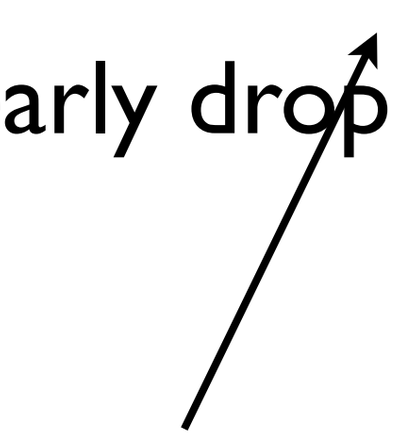


TBB Structures

- TBB structure is a combination of TBBs

U need an iphone lol ==> RT @UserB: @UserA

i nearly dropped my blackberry in that poool :(



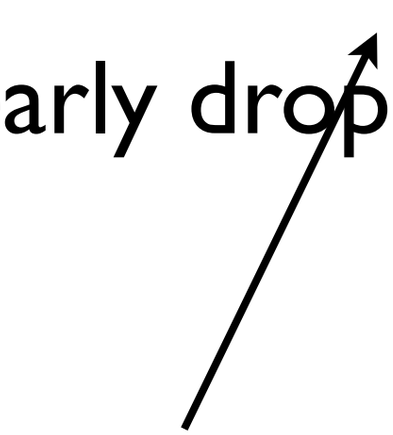
COM

TBB Structures

- TBB structure is a combination of TBBs

U need an iphone lol ==> RT @UserB: @UserA

i nearly dropped my blackberry in that poool :(

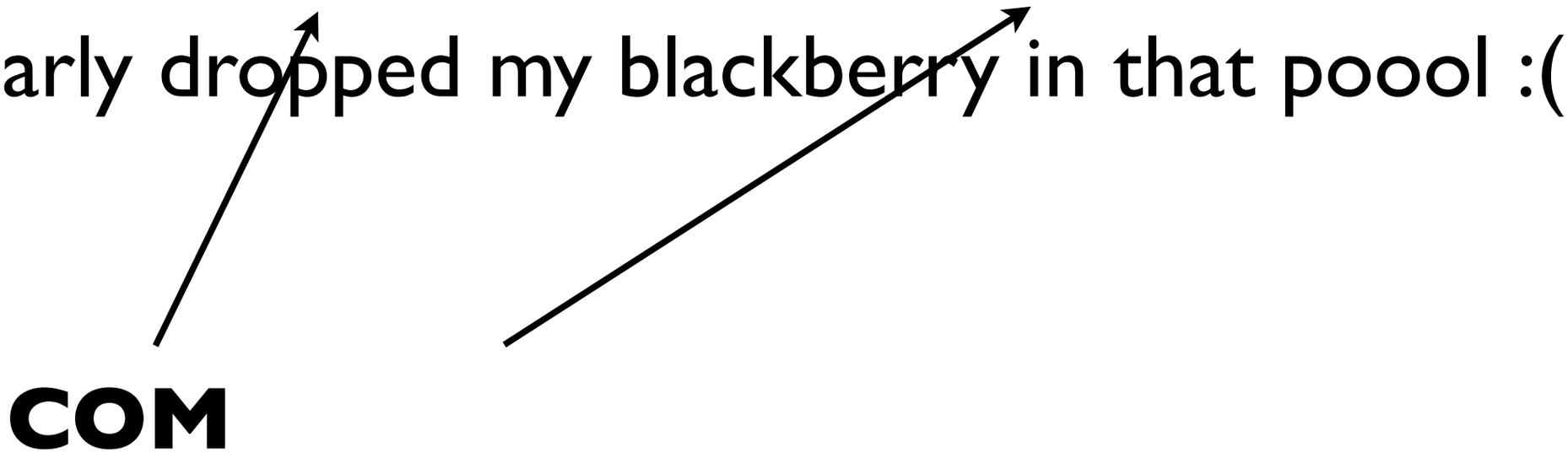


TBB Structures

- TBB structure is a combination of TBBs

U need an iphone lol ==> RT @UserB: @UserA

i nearly dropped my blackberry in that poool :(

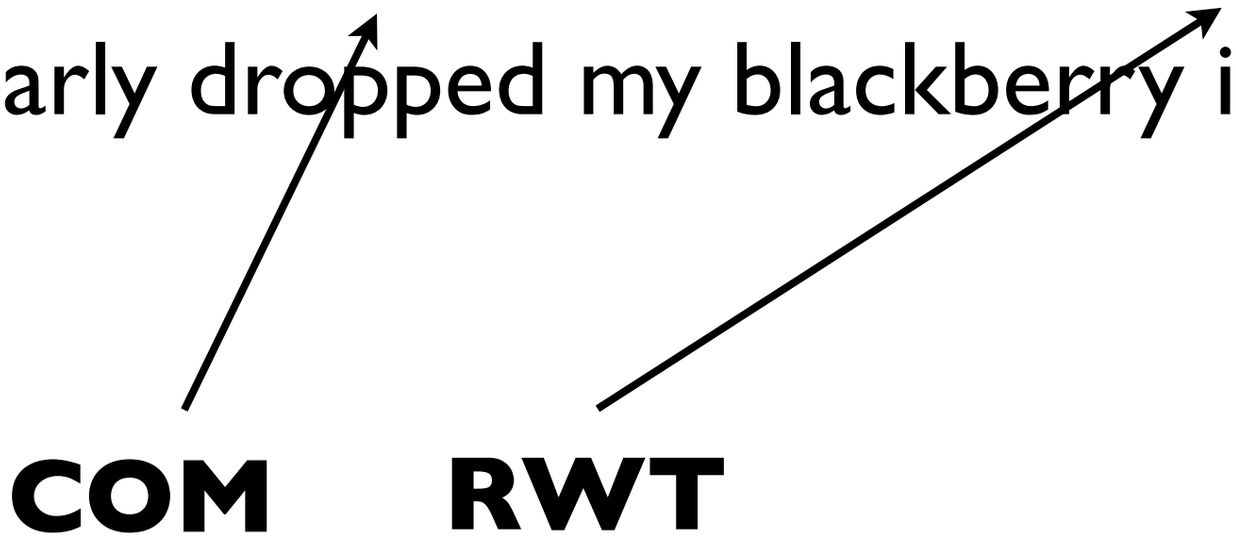


TBB Structures

- TBB structure is a combination of TBBs

U need an iphone lol ==> RT @UserB: @UserA

i nearly dropped my blackberry in that poool :(

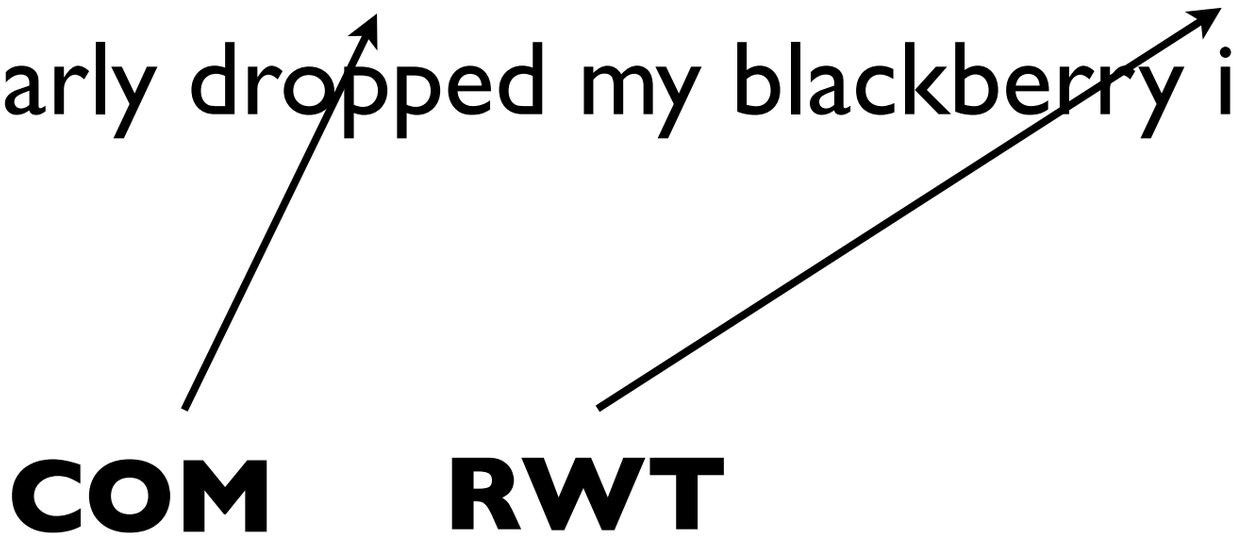


TBB Structures

- TBB structure is a combination of TBBs

U need an iphone lol ==> RT @UserB: @UserA

i nearly dropped my blackberry in that poool :(

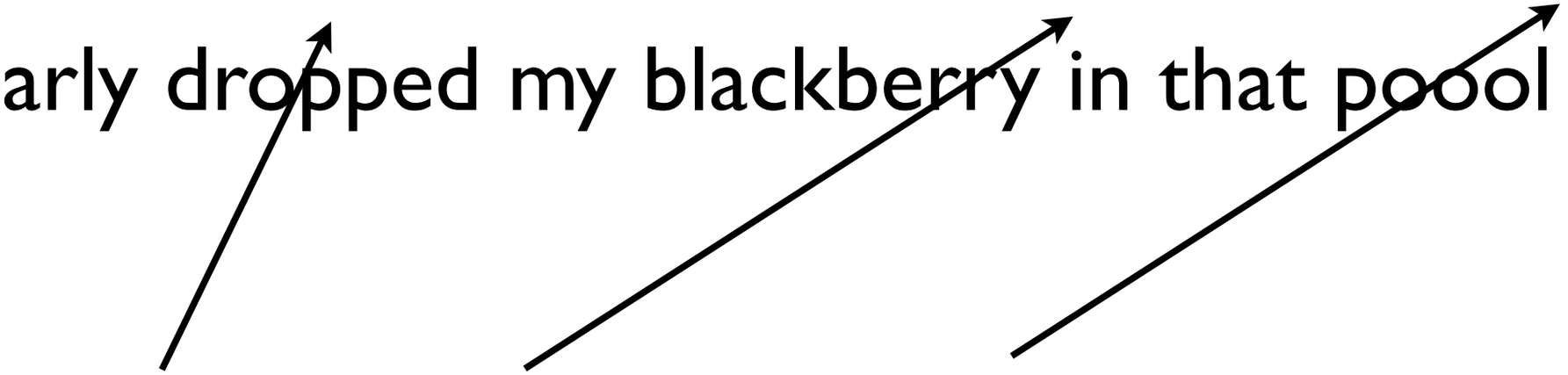


TBB Structures

- TBB structure is a combination of TBBs

U need an iphone lol ==> RT @UserB: @UserA

i nearly dropped my blackberry in that poool :(

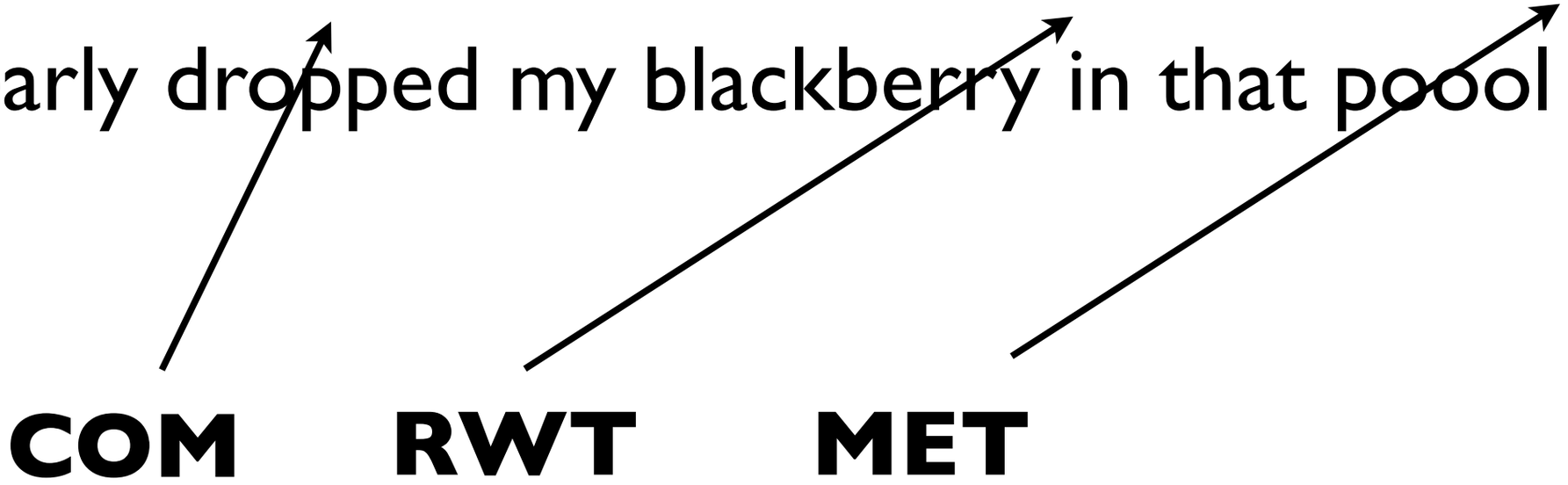


TBB Structures

- TBB structure is a combination of TBBs

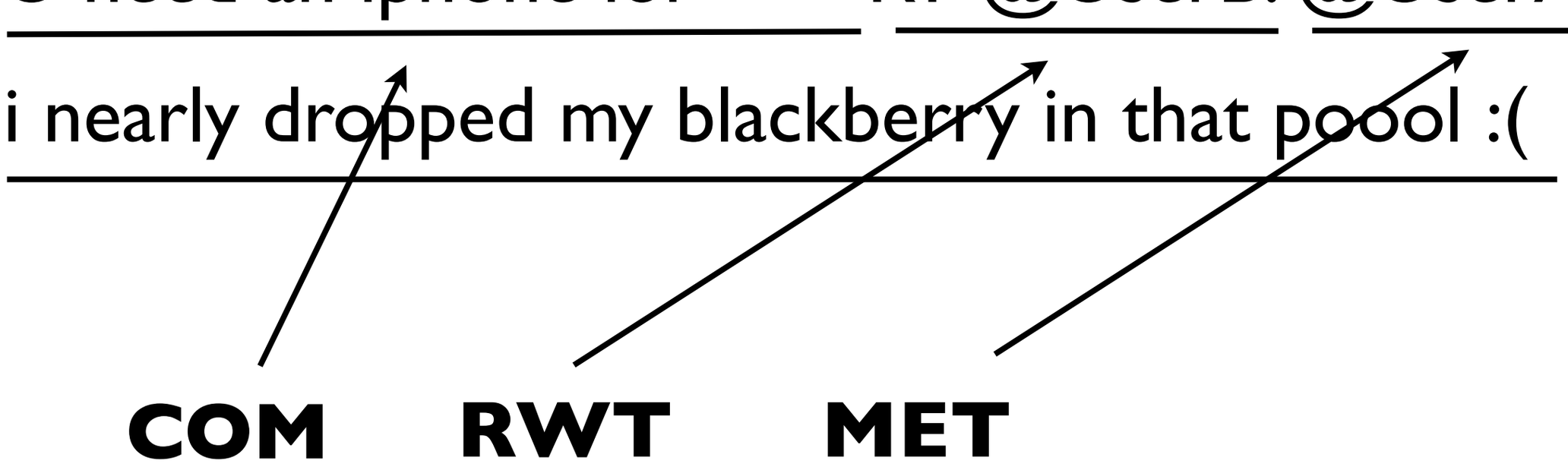
U need an iphone lol ==> RT @UserB: @UserA

i nearly dropped my blackberry in that poool :(



TBB Structures

- TBB structure is a combination of TBBs

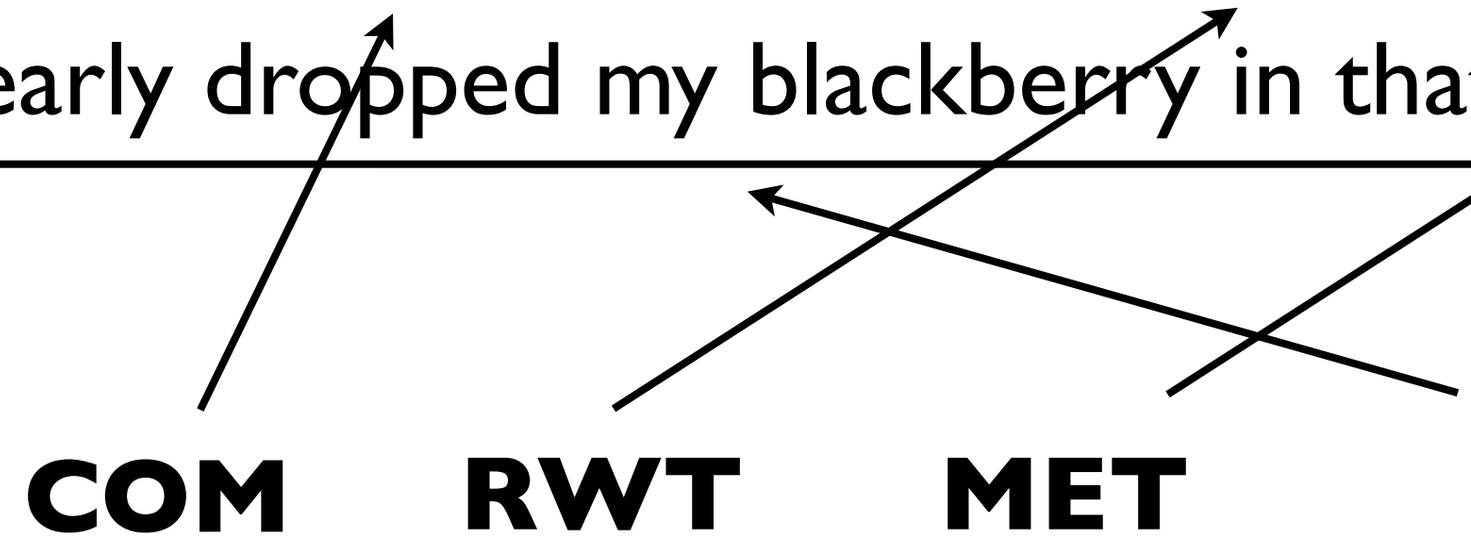


TBB Structures

- TBB structure is a combination of TBBs

U need an iphone lol ==> RT @UserB: @UserA

i nearly dropped my blackberry in that poool :(

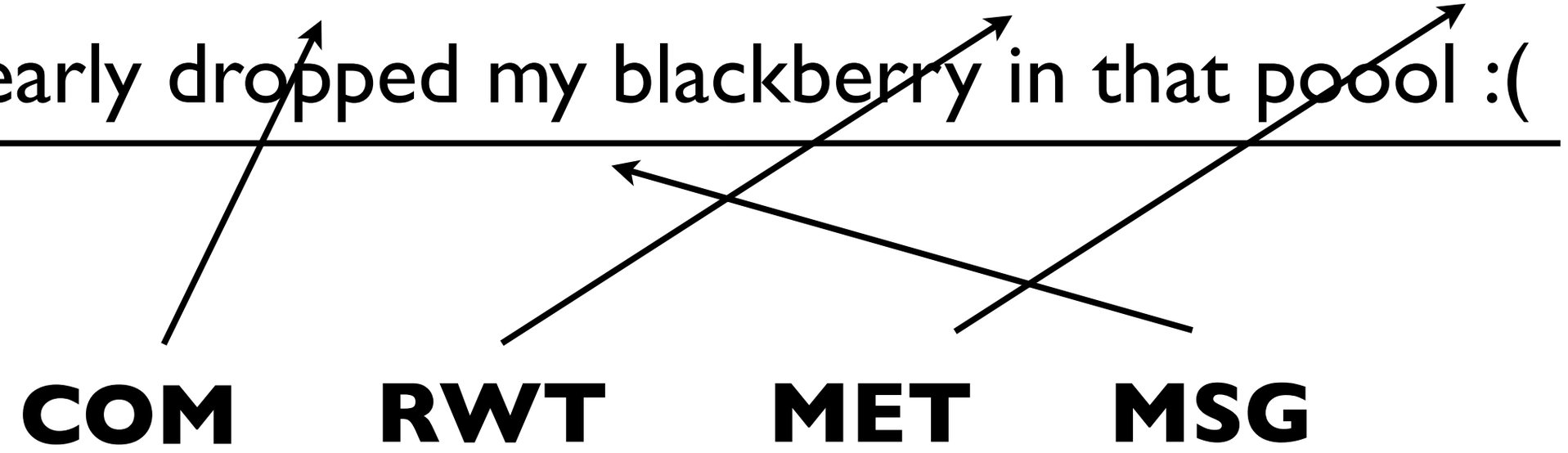


TBB Structures

- TBB structure is a combination of TBBs

U need an iphone lol ==> RT @UserB: @UserA

i nearly dropped my blackberry in that poool :(



TBB Structures

- TBB structure is a combination of TBBs

U need an iphone lol ==> RT @UserB: @UserA

i nearly dropped my blackberry in that poool :(

COM **RWT** **MET** **MSG**

- TBB Structure is "**COM RWT MET MSG**".

TBB Structures

- TBB structure is a combinations of TBB

TBB Structures

- TBB structure is a combinations of TBB

  New IPhone in Semptember ----

  http: //buswk.co/jbyCo #iphone #apple

TBB Structures

- TBB structure is a combinations of TBB

  New IPhone in Semptember ----

  http: //buswk.co/jbyCo #iphone #apple

TBB Structures

- TBB structure is a combinations of TBB

New IPhone in Semptember ----

http: //buswk.co/jbyCo #iphone #apple

TBB Structures

- TBB structure is a combinations of TBB

New IPhone in Semptember ----

http: //buswk.co/jbyCo #iphone #apple

MSG

TBB Structures

- TBB structure is a combinations of TBB

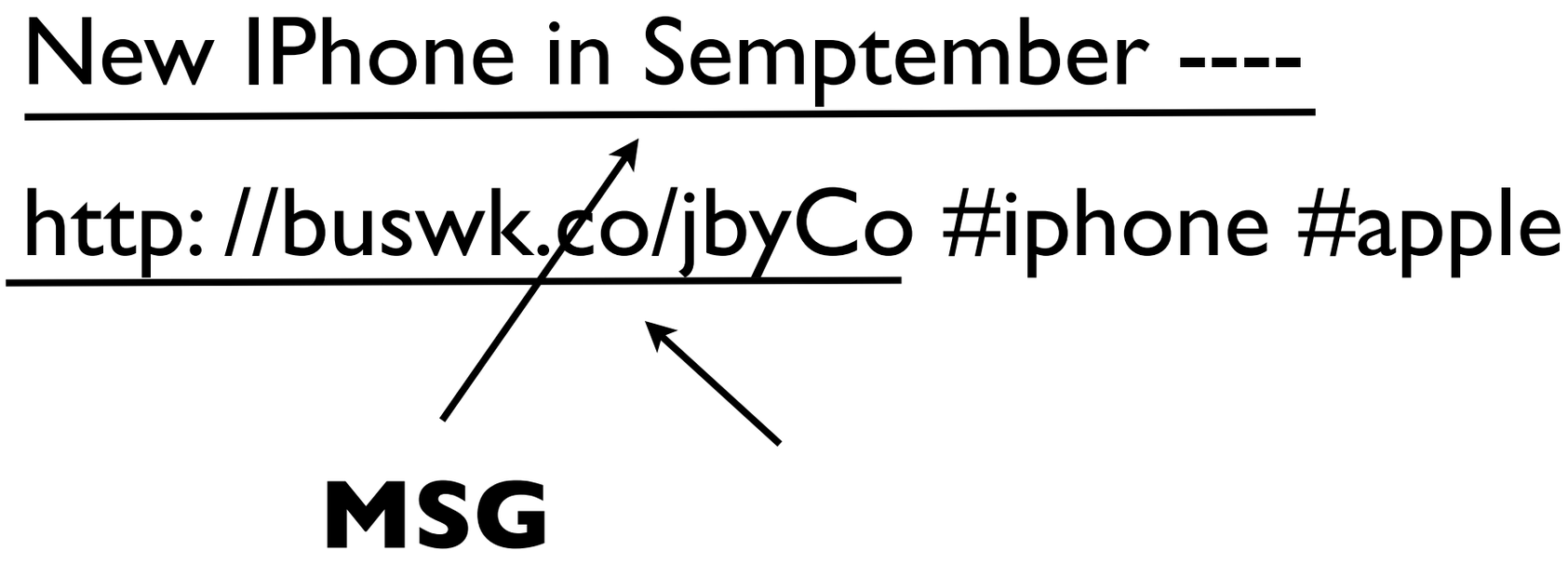
New IPhone in Semptember ----

http: //buswk.co/jbyCo #iphone #apple

MSG

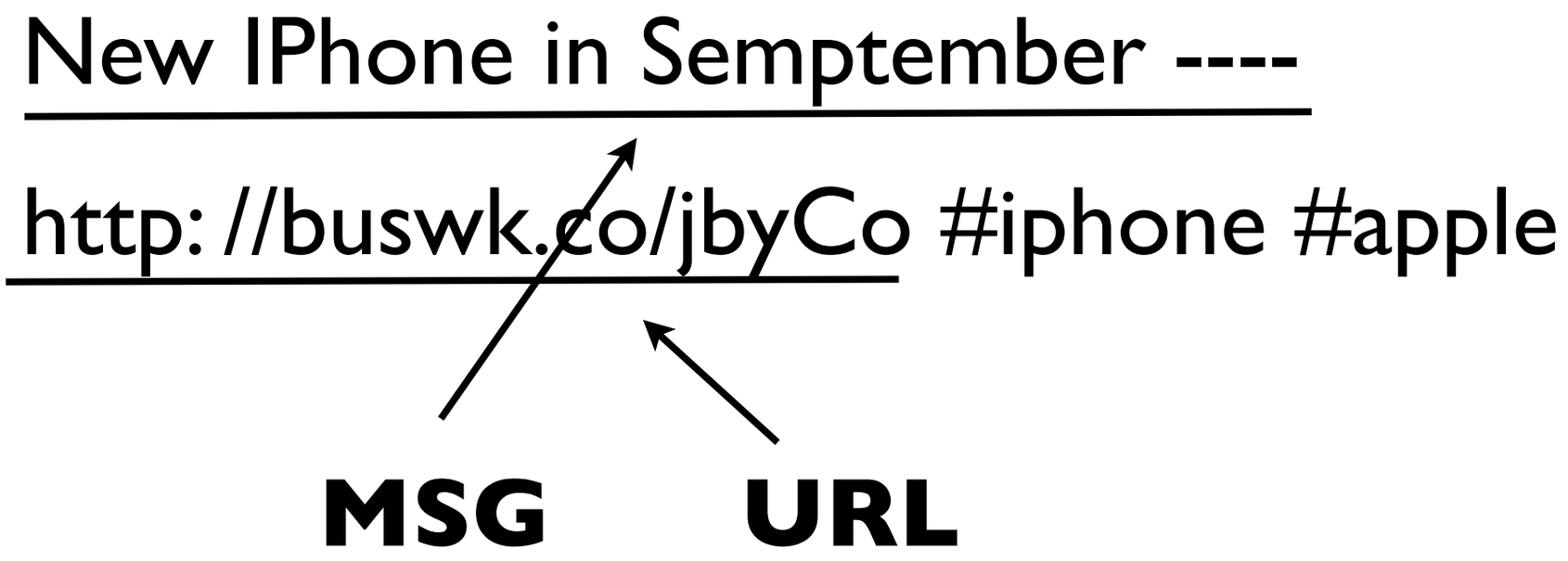
TBB Structures

- TBB structure is a combinations of TBB



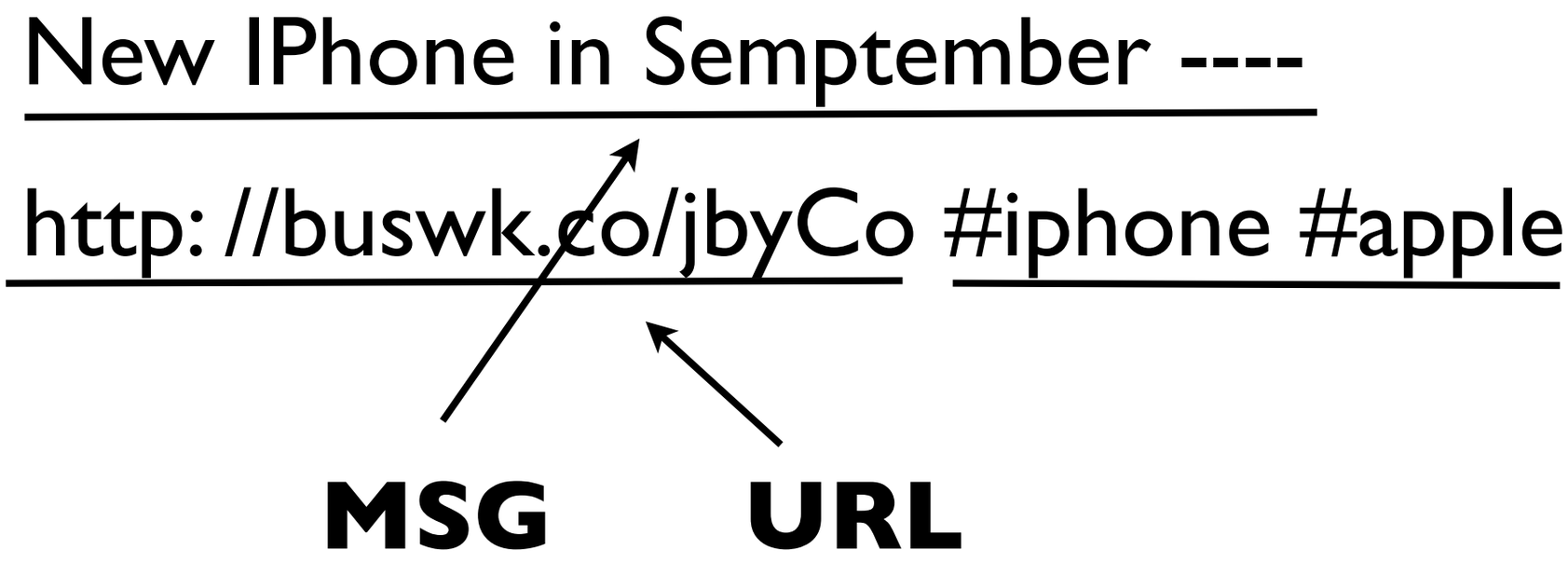
TBB Structures

- TBB structure is a combinations of TBB



TBB Structures

- TBB structure is a combinations of TBB



TBB Structures

- TBB structure is a combinations of TBB

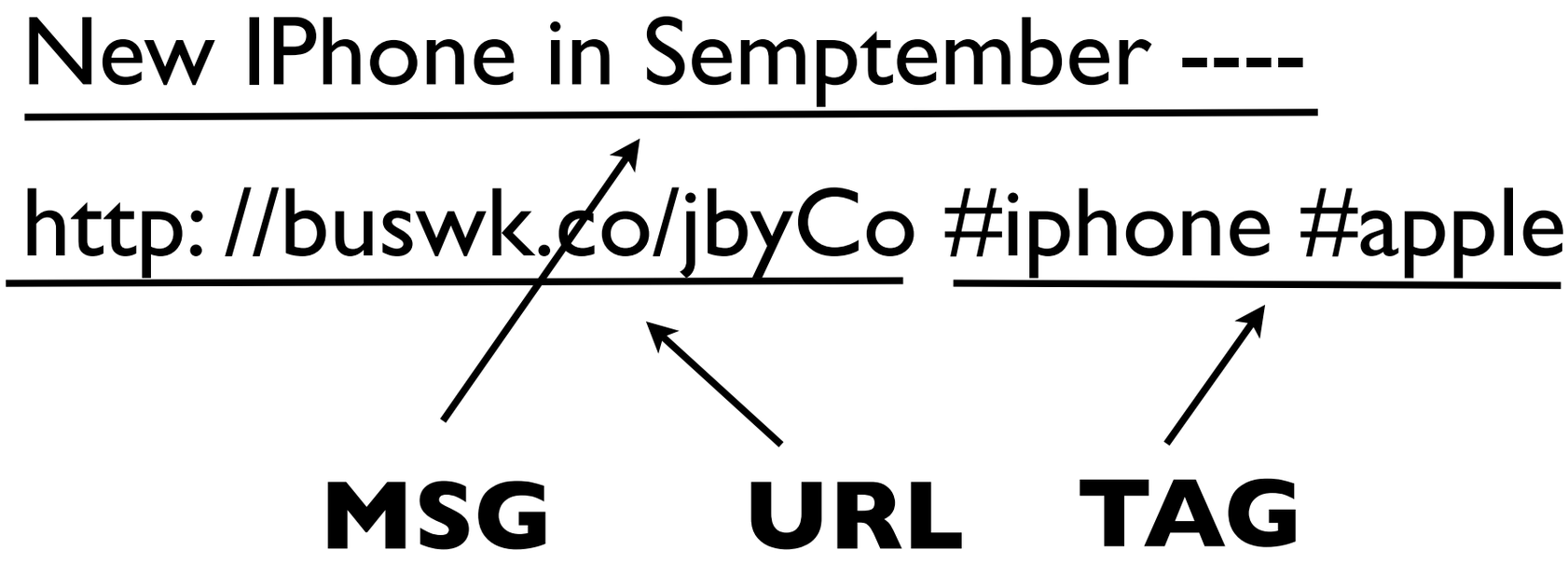
New IPhone in Semptember ----

http: //buswk.co/jbyCo #iphone #apple

MSG **URL**

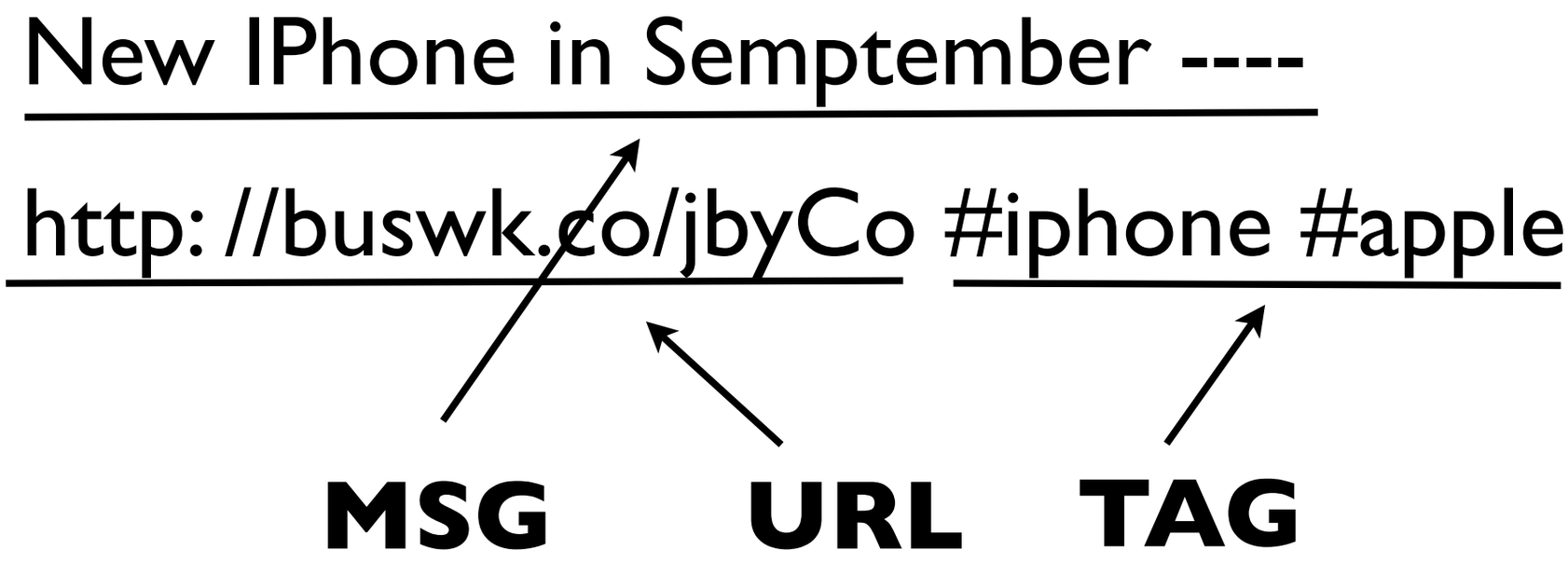
TBB Structures

- TBB structure is a combinations of TBB



TBB Structures

- TBB structure is a combinations of TBB



- TBB Structure is "**MSG URL TAG**".

TBB Structures Distribution

TBB Structures Distribution

- 14 most frequent TBB Structures in Twitter.
- "OTHERS" accounts for all other TBB Structures.

TBB Structures Distribution

- 14 most frequent TBB Structures in Twitter.
- "OTHERS" accounts for all other TBB Structures.

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		

TBB Structures Distribution

- 14 most frequent TBB Structures in Twitter.
- "OTHERS" accounts for all other TBB Structures.

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		

- People use simple and fixed structures to tweet.

Automatic TBB Tagger

- Sequence labeling approach (Conditional Random Field).
- Features for TBB tagger:
  - Token type; Pos; Length; Prefix and suffix; Twitter orthography (e.g, the preceding of "RWT" is more likely to be "COM").
- TBB structure identification achieves an accuracy of 82.60%.
  - #(Train dataset)=1000; #(Dev dataset)=500; #(Test dataset)=500;

TBB Analysis

TBB Analysis

- Clustering tweets by TBB structures.

TBB Analysis

- Clustering tweets by TBB structures.
- Each cluster has similar characteristics:

TBB Analysis

- Clustering tweets by TBB structures.
- Each cluster has similar characteristics:
  - Public Broadcast: MSG URL; MSG URL TAG
    - E.g., Apple brings new iPad to China http://bbc.in/NI2HW9

TBB Analysis

- Clustering tweets by TBB structures.
- Each cluster has similar characteristics:
  - Public Broadcast: MSG URL; MSG URL TAG
    - 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-12)
    - E.g, *I thought we were isolated and no one would want to invest here! RT @UserA: Honda announces 500 new jobs in Swindon*

TBB Analysis

- Clustering tweets by TBB structures.
- Each cluster has similar characteristics:
  - Public Broadcast: MSG URL; MSG URL TAG
    - 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-12)
    - 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
    - E.g, RT @UserA: Forreal doeee? (Wanda voic) #Icant cut it out #Newark http://twipic.com/2u15xa...lmao!!WOW ... http://tmi.me/

TBB Analysis: OOV

- Out-of-Vocabulary Value for TBB Structures:

TBB Structures	O.(%)	TBB Structures	O.(%)
OTHERS	**4.30**	MET MSG URL	1.42
TAG MSG URL	3.42	MSG	1.32
MSG URL	1.93	MSG TAG	1.31
MSG URL TAG	1.91	RWT MSG URL	1.30
COM RWT MSG	1.80	MET MSG	1.15
TAG MSG URL	1.78	**RWT MSG**	**0.82**
MSG MET MSG	1.64	**RWT MSG TAG**	**0.58**
TAG MSG	1.63		

- People retweet high quality text.
- More blocks = More OOV words.

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 the corresponding TBB.
  - The context information of the TBB containing the query.
  - The number of blocks in a tweet.

Rank Approaches

- Baseline (Duan et al,Coling-10):
  - Features: Length; BM25 score; Link (the most effective feature).
- SM_Rank:
  - Features: More social media features (e.g, number of followers).
- TBB_Rank:
  - Features: Our TBB features.

Experiment

- Dataset: 100 queries and 936 judged tweets
- Leaning to rank model: SVM^{Rank}
- Evaluation: Mean Average Precision (MAP)
- Ten-fold cross-validation

Experimental Result

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

- TBB is effective for Twitter Retrieval!

The Most Important TBB Structure for Twitter Retrieval

- Link feature is the most important feature in Baseline (Duan et al,Coling-10).
- Replace the Link feature by TBB Structure Type feature related "URL" block in Baseline.
- "MSG URL" is the most important structure!

	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.

Thanks!