What Are Tweeters Doing: Recognizing Speech Acts in Twitter

Renxian Zhang ***Dehong Gao*** ***Wenjie Li***

Department of Computing, The Hong Kong Polytechnic University, Hong Kong

Introduction

Speech Acts in Twitter

- Informing, requesting, expressing sentiment, etc. (Austin, 1962)
- Due to its social networking nature, twittering is a communicative act with explicit intentions.

Our Types	Searle's (1975) Types
Statement	Assertive
Question	Directive
Suggestion	
Comment	Expressive
Miscellaneous	Commissive
	Declarative

Significance of Recognizing Speech Acts in Twitter

- For Twitter itself, it helps to reveal how a topic is constituted in terms of tweeters' speech acts and whether there is any topic shift.
- For tweet posters, it helps us to understand their behavior as a community defined by their common interest in a certain topic, as well as their behavior as individuals susceptible to the others in the same community.
- For tweet readers, the distribution of speech act types under a specific topic provides information to help them become efficient readers in a sea of tweets.

Research Agenda

- Treating Twitter speech act recognition as a multi-class classification problem
- Finding robust features in the face of extreme noisiness (spelling errors, grammatical mistakes, etc.) of Twitter data
- Exploring on what level training data should be prepared – topic-level, category-level, or the whole Twitter space – and whether one model trained on one topic/category can adapt to a different topic/category, thus helping to define the scope of the task and inform training data preparation for the task

Textual Features

Word-based Features

- Cue-words and phrases
  - The lexical limitation and Twitter's deviation from standard English defies the use of manually crafted lexicons.
  - Collecting 531 high-frequency unigrams, bigrams, and trigrams and manual checking

	Examples	Total
Unigrams	*know, hurray, omg, pls, why*	268
Bigrams	*do it, i bet, ima need, you can*	164
Trigrams	*?!?, heart goes out, rt if you*	99

- Non-cue words
  - Abbreviations and acronyms

    1 binary feature for each of the 1153 tokens (e.g., *4ever*, *tq*) collected from online and published resources
  - Opinion words

    1 binary feature for each of the 2460 tokens (e.g., *shallow*, *vague*) collected from the SentiWordNet and Wilson Lexicon
  - Vulgar words

    1 binary feature for each of the 341 tokens (e.g., *c**t*, *f**k*) collected from an online resource
  - Emoticons

    1 binary feature for each of the 276 tokens (e.g., O:), *-*) collected from an online resource

Character-based Features

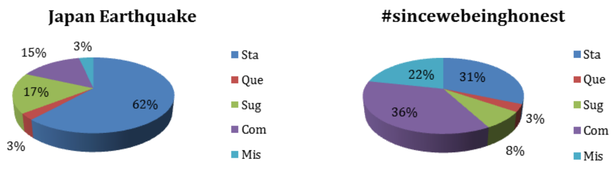
- Twitter-specific symbols
  - # (statements, comments), @ (questions, suggestions), RT (statements)
  - 1 binary-valued feature for initial position of each and 1 ternary-valued feature for frequencies of each (0, 1-2, 2+)
- Indicative punctuations
  - ? (questions), ! (comments, suggestions)
  - 1 ternary-valued feature for frequencies of each (0, 1-2, 2+)

Experiments

Data

Category	Topic	# Tweets
News	Japan Earthquake (JE)	1742
	Libya Releases (LR)	1408
Entity	Dallas Lovato (DL)	677
	Nikki Taylor (NT)	786
LST	#100factsaboutme (FM)	2000
	#sincewebeinghonest (SH)	2000

- Collected from Twitter.com from March 1, 2011 to March 31, 2011
- Examples of speech act distribution



Results

- Using different features

Feature	Sta	Que	Sug	Com	Mis	Avg
Cue	0.788	0.455	0.554	0.623	0.422	0.668
Non-cue	0.671	0.088	0.068	0.355	0.074	0.447
Symbols	0.681	0.473	0.039	0.412	0.097	0.483
All	0.798	0.597	0.564	0.670	0.446	0.695

- Conclusions
  - Cue words and phrases are most valuable features and the character-based symbols are very useful.
  - Mixed training/test is nearly as successful as per-category or per-topic training/test.
  - Heterogeneous training/test configuration is feasible.

- Using different training/test configuration

	Avg
JE : JE	0.749
LR : LR	0.891
DL : DL	0.779
NT : NT	0.649
FM : FM	0.664
SH : SH	0.533
All	0.695

	Avg
News : News	0.810
Entity : Entity	0.716
LST : LST	0.549
All	0.673

	Avg
All : All	0.639

same training: test

	Avg
LR : JE	0.575
JE : LR	0.792
NT : DL	0.414
DL : NT	0.425
SH : FM	0.592
FM : SH	0.386

	Avg
Entity : News	0.594
LST : News	0.167
News : Entity	0.507
LST : Entity	0.420
News : LST	0.344
Entity : LST	0.260

different training: test