Automatic Twitter Topic Summarization with Speech Acts

Renxian Zhang, Wenjie Li, Dehong Gao, and You Ouyang[1]

***Abstract*—With the growth of the social media service of Twitter, automatic summarization of Twitter messages (tweets) is in urgent need for efficient processing of the massive tweeted information. Unlike multi-document summarization in general, Twitter topic summarization must handle the numerous, short, dissimilar, and noisy nature of tweets. To address this challenge, we propose a novel speech act-guided summarization approach in this work. Speech acts characterize tweeters' communicative behavior and provide an organized view of their messages. Speech act recognition is a multi-class classification problem, which we solve by using word-based and symbol-based features that capture both the linguistic features of speech acts and the particularities of Twitter text. The recognized speech acts in tweets are then used to direct the extraction of key words and phrases to fill in templates designed for speech acts. Leveraging high-ranking words and phrases as well as topic information for major speech acts, we propose a round-robin algorithm to generate template-based summaries. Different from the extractive method adopted in most previous works, our summarization method is abstractive. Evaluated on two 100-topic datasets, the summaries generated by our method outperform two**

Manuscript submitted on March 21, 2012. The work described in this paper was supported by the grant GRF PolyU 5230/08E.

Renxian, Zhang, Wenjie, Li, and Dehong Gao are with the Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong (e-mail: csrzhang@ comp.polyu.edu.hk).

You Ouyang was with the Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong. He is now with Miaozhen Systems, Beijing, China.

kinds of representative extractive summaries and rival human-written summaries in terms of explanatoriness and informativeness.

***Index Terms*—Twitter; speech act; abstractive summarization; key word/phrase extraction**

I. INTRODUCTION

In the age of social media, the problem of information overload is scaling up at an unprecedented rate as massive information inundates information consumers. According to a *Wall Street Journal* report[2], the microblogging service of Twitter spews out over 200 million tweets every day. The top trending topics on *Twitter.com* each comprises thousands of tweets or more, deterring attempts to read all the tweets under a topic. A promising solution lies in text summarization techniques, which can generate a synopsis of a mass of tweets under a topic with information distilled from them. Summarizing Twitter topics, however, is a very different challenge from summarizing other genres of text such as news articles, research papers, books, etc.

[2] http://on.wsj.com/r8bLkn

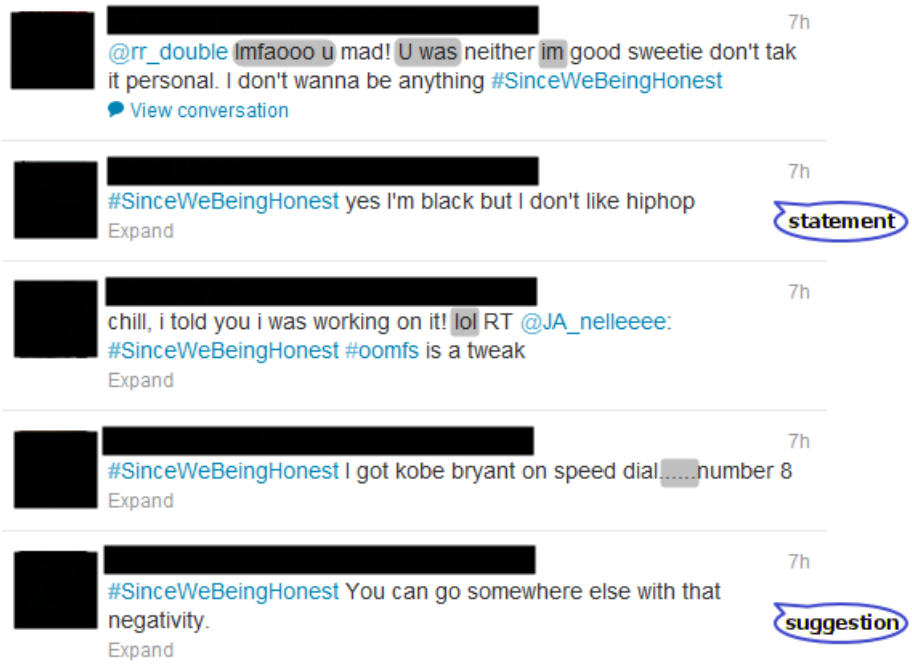


Fig. 1. A snapshot of *#sincewebeinghonest* tweets with our annotations.

Let's explicate the major differences with the aid of a snapshot (Fig. 1) of several tweets under the topic of *#sincewebeinghonest*. We blotched the user accounts and icons to hide people's identities and annotated the tweets to facilitate our explanations in the following.

1) By nature, Twitter topic summarization is a kind of multi-document summarization. Typical multi-document summarization tasks, such as those for newswire, deal with dozens of documents each with several hundred words or several dozen sentences[3]. By contrast, the tweets under a given topic usually number in the thousands, or tens of thousands, with each tweet being no more than 140 characters long. As is shown in Fig. 1, a tweet consists of only one or two sentences.

2) Typical multi-document summarization tasks are targeted at closely related documents, such as news reports about the same event that overlap considerably in their contents[4]. By contrast, tweets under a topic are only loosely lumped together, sharing not much in common. The tweets in Fig. 1, for

[3] http://www-nlpir.nist.gov/projects/duc/index.html

[4] See for an example http://newsblaster.cs.columbia.edu/summaries/2012-02-14-03-22-49-006.html

example, make a miscellany of personal affairs, black people, music, sports star, and chitchat, though all belonging to #*sincewebeinghonest*.

3) The source texts to be summarized are usually formal writings and the language quality is high. By contrast, the language of tweets is highly noisy, rife with nonstandard usage, spelling and grammar mistakes, mixed symbols and characters, Netspeak expressions, etc. Some examples are ……, *lol*, *Imfaooo*, *U was*, *im*, which are shaded in Fig. 1.

Due to the above characteristics, text summarization techniques in general may not adapt well to the Twitter text. To take the challenge, in this paper we will develop an approach to Twitter topic summarization, which is designed to overcome the difficulties caused by the Twitter idiosyncrasies.

The most original part of our approach is the use of **speech acts**, which capture the common grounds of tweets from a communicative perspective. When communicating with tweets, users may share information, ask questions, make suggestions, express sentiments, etc. which are all instances of "speech acts" [1]. Each tweet is associated with a type of speech act, like the "statement" and "suggestion" for the two tweets in Fig. 1.

Unless in a few cases (e.g., using a speech act hashtag like #*question*), users do not report the speech acts they are performing when twittering. So before using speech acts for summarization, we need to recognize them in tweets automatically. Then, guided by the recognized speech acts in the tweets, we can proceed to extract key words and phrases from the tweets. Leveraging the linguistic knowledge of speech acts, we generate summaries that integrate the extracted language materials into speech act-based sentence templates.

Before presenting the technical details of our work, we would like to point out that our approach is particularly suitable for Twitter topic summarization because it enables us 1) to deal with a few

clusters of communicatively similar tweets, each cluster being a speech act type, instead of a large medley of tweets; 2) to establish connections between seemingly unrelated words and expressions; 3) to resist noisy text in summaries by key word/phrase extraction. The major contributions of our work are summarized in the following.

- We propose a **speech act-based approach** to Twitter Topic summarization. Most existent Twitter summarization methods follow the frameworks of general text summarization.
- We produce **abstractive summaries**, which fit the numerous, short, and jumbled nature of tweets. Most existent Twitter summarization methods are extractive.
- We arrive at interesting findings about noise in Twitter text. For our task at the least, intensive and expensive text de-noising or **normalization can be avoided**.

The remainder of the paper is organized as follows. Section II reviews related work in speech act recognition and Twitter summarization. Sections III to V unfold the technical details of the three core modules of our approach: recognizing speech acts in tweets, extracting speech act-guided key words/phrases, and generating abstractive summaries. Section VI presents extensive experimental results of evaluating the summaries. Section VII concludes the work and discusses future work.

II. RELATED WORK

The current work hinges on the important notion of speech act, which was proposed half a century ago by Austin [1] and the speech act theory has since established itself in pragmatics. Over the years, linguists have shown sustained interest in the taxonomical [2] and logical aspects [3] of speech acts.

In computational linguistics, speech act is also extended to dialogue act [4] to accommodate more conversational phenomena such as grounding or turn-taking. For both speech act and dialogue act, the main interest is in their automatic recognition to model conversation [5] [6], which relies on

annotated corpora such as Switchboard-DAMSL [7] and Meeting Recorder Dialog Act [8]. Prior to the flourish of microblogging services such as Twitter, speech act recognition has been extended to electronic media such as email and discussion forum [9] [10] in order to study the behavior of email or message senders.

But neither the corpora for ordinary verbal communications nor the methods developed for email or discussion forum can be directly used for speech act recognition in Twitter, a new classification task started only recently [11] and hampered by the deficiency of annotated data. Semi-supervised [12] and unsupervised methods [13] can help to alleviate this problem.

What further complicates this task is a distinctive Netspeak style that is situated between speech and text but resembles neither [14] [15] and the notorious noisiness of the text. Many researchers believe that Twitter text normalization [16] [17] is necessary for various NLP tasks.

Automatic summarization on Twitter or microblogs in general is a special kind of multi-document summarization. An early successful multi-document summarizer is SUMMON [18]. Present-day representative approaches include the centroid-based model [19], graph-based model [20], and clustering-based model [21].

Among the several published works on Twitter summarization, Sharifi et al. [22] find important phrases to be included in a summary with a graph-based algorithm, but the authors later [23] develop a simpler “Hybrid TF-IDF” method, which ranks tweet sentences using the TF-IDF scheme and produces even better results. This is also confirmed by Inouye [24], who shows that Hybrid TF-IDF outperforms several other mainstream summarization approaches, including MEAD [19], LexRank [25], and TextRank [20]. A more complicated work is reported by Liu et al. [26], which highlights the use of linked webpage content and relies on Integer Linear Programming-based optimization to

extract tweet sentences.

The above efforts are all adaptations of extractive summarization methods on text of other domains to the Twitter text. By contrast, we propose an abstractive approach to summarizing Twitter topics based on template-based Natural Language Engineering (NLG). This is a well-understood area [27] with many practical systems (e.g., [28] [29]). It has been used to generate abstractive summaries for news articles [18], technical articles [30], evaluative text [31], and briefings [32], but not microblogs like Twitter to the best of our knowledge.

III. Twitter Speech Act Recognition

In this section, we present our work on speech act recognition (**SAR** hereafter) for Twitter text, as a prerequisite for speech act-guided key word/phrase extraction and summarization.

A. Types of Speech Acts in Twitter

The scope of Twitter SAR is based on Searle's [2] popular taxonomy of speech acts: assertives (asserting something's being the case), commissives (committing the speaker to some future action), directives (getting the hearer to do something), declaratives (bringing about a different state of world by uttering something), and expressives (expressing the speaker's psychological state).

Table I lists all the 5 speech act types we use, alongside the corresponding Searle's types and examples from our experimental datasets. A tweet belongs to one of 4 genuine types of speech act – **statement**, **question**, **suggestion**, **comment** – or the **miscellaneous** type. Our choice stems from the fact that unlike face-to-face communication, twittering is more in a broadcasting style than on a personal basis. Statement and comment correspond to Searle's assertives and expressives, which are usually intended to make one's knowledge, thought, and sentiment known. Searle's directives correspond to our question and suggestion, which are distinct speech acts targeted at other tweeters.

Both commissives and declaratives are rare, as are other interpersonal speech acts such as "threat" or "thank". So they are all relegated to "miscellaneous".

TABLE I
SEARLE'S [2] SPEECH ACT TYPES, OUR SPEECH ACT TYPES AND EXAMPLES

Searle's Types	Our Types	Example Tweets
Assertive	Statement	*Libya Releases 4 Times Journalists - http://www.photozz.com/?104k*
Directive	Question	*#sincewebeinghonest why u so obsessed with what me n her do?? Don't u got ya own man???? Oh wait.....*
	Suggestion	*RT @NaonkaMixon: I will donate 10 $ to the Red Cross Japan Earthquake fund for every person that retweets this! #PRAYFORJAPAN*
Expressive	Comment	*is enjoying this new season of #CelebrityApprentice.... Nikki Taylor = Yum!!*
Commissive	Miscellaneous	*65. I want to get married to someone i meet in highschool. #100factsaboutme*
Declarative		

Assuming one tweet demonstrates only one speech act type, Twitter SAR is a five-class single-label classification problem. It is possible that one tweet demonstrates more than one speech act type. But given the short length of tweets, multi-speech act tweets are rare and our simplifying assumption is effective in reducing the complexity of the problem.

B. Feature Set Design

In the following, we describe the feature sets used for recognizing the five types of speech act in Table I, including word-based and symbol-based features.

1) Word-based Features

We have two major types of 535 word-based features, all of which are binary-valued.

- **Cue Words and Phrases**

Some speech acts are typically signaled by some cue words or phrases, such as *whether* for "question" and *could you please* for "suggestion". There are some manually compiled lexicons for speech act cues (e.g., [33]), but we refrain from using them for two reasons. First, the cue lexicons are

very limited, consisting mostly of verbs. But words of other part of speech (including closed-class words) and phrases may be equally predictive. Second, such lexicons only serve standard English, not Twitter English rife with non-standard spellings, acronyms, and abbreviations. Therefore, we manually compiled a speech act cue lexicon of Twitter English from a dataset of 10K tweets, which are not used in the experiments. First, high-frequency unigrams, bigrams, and trigrams are collected for "statement", "question", "suggestion", and "comment". Then we employed a linguistics student to manually check them and come up with a total of 531 such features. Table II shows some examples.

TABLE II
EXAMPLES OF CUE WORDS AND PHRASES

	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**

Some special words, though not intuitively cuing speech acts, may indirectly signal speech acts. We use four types of such non-cue words explained in the following.

Abbreviations and Acronyms: one feature indicates whether such shortened word forms appear. We collected the lexicon from online[5] and published [14] resources, a total of 1153 words. Examples are *4ever* for "forever" and *tq* for "thank you". We then restore the shortened words to their original forms before extracting the next two features: opinion words and vulgar words.

Opinion Words: one feature indicates whether opinion words appear. To judge opinion words, we used the SentiWordNet [34] and Wilson Lexicon [35] widely used for opinion mining. As we are only interested in strong opinion words, we build a lexicon by intersecting highly opinionated words (positive score + negative score $\geq$ 0.5, note that both positive score and negative score are non-negative) from the SentiWordNet with the "strong" words from the Wilson Lexicon, obtaining

[5] http://www.chatslang.com

2460 words, like *shallow*, *vague*, *scary*, etc.

Vulgar Words: one feature indicates whether vulgar words appear. We used the API from an online resource[6] and collected 341 such words as *c**t* and *f**k*[7].

Emoticons: one feature indicates whether emoticons appear. We collected 276 emoticons from an online resource[8], such as O:) and *-*.

2) Symbol-based Features

We have two types of eight symbol-based features, which indicate the frequency and position of special characters and are either binary- or ternary- valued.

- **Twitter-specific Symbols**

We concentrate on the three symbols specific to Twitter: *#*, *@*, and *RT*. # is a hashtag marker often used in a mention of something to be stated about or commented on; @ is a prefix to a tweeter account, which tends to be associated with the more interpersonal speech acts of questions or suggestions; *RT* stands for "retweet" and its presence, especially in the initial position, strongly indicates a statement. Repeated use of them is an even stronger indicator of possible speech act types. Each of those symbols is associated with two features: one binary-valued feature indicating whether the symbol is in the initial position of a tweet and one ternary-valued feature indicating whether the symbol does not appear (0), appears one or two times (1), or appears more than two times (2).

- **Indicative Punctuations**

We single out two punctuations: *?* and *!* as the former often indicates a question and the latter is likely to indicate a comment or suggestion. Each of them is associated with 1 ternary-valued feature indicating zero appearance (0), one or two appearances (1), or three or more appearances (2).

[6] http://www.noswearing.com/dictionary

[7] For ethical concerns, we mask part of the words here and deliberately avoid using them in other examples.

[8] http://www.sharpened.net/emoticons/

C. Classification Evaluation

We evaluated our feature sets on 6 Twitter datasets with hand-annotated speech act types as labels, using SVM with a linear kernel as our classifier.

1) Data Preparation

Using the Twitter search API, we collected tweets of 6 randomly chosen trending topics on *Twitter.com* from March 1, 2011 to March 31, 2011. The topics fall into three categories – News, Entity, Long-standing Topic (LST) – that correspond to the three "topic types" [36]. We manually annotated all the 8613 tweets as one of Sta (statement), Que (question), Sug (suggestion), Com (comment), or Mis (Miscellaneous)[9]. The categories, topics and tweet numbers are shown in Table III.

TABLE III
DETAILS OF EXPERIMENTAL DATASETS

Category	Topic	# Tweets
News	Japan Earthquake	1742
	Libya Releases	1408
Entity	Dallas Lovato	677
	Nikki Taylor	786
LST	#100factsaboutme	2000
	#sincewebeinghonest	2000

Different categories/topics of tweets have different speech act distributions. Fig. 2 illustrates the speech act distributions in all the 6 topics we used. Obviously, statements and comments take the majority. Generally speaking, entity topics are dominated by comments and news topics by statements. Special cases also exist, such as "Japan Earthquake" containing a considerable proportion of suggestions (e.g., about what people can do to help victims). The imbalanced distribution of speech act types in Twitter topics will be a crucial evidence for designing our summarization algorithm.

[9] www4.comp.polyu.edu.hk/~csrzhang/files/Public%20datasets.tar.gz

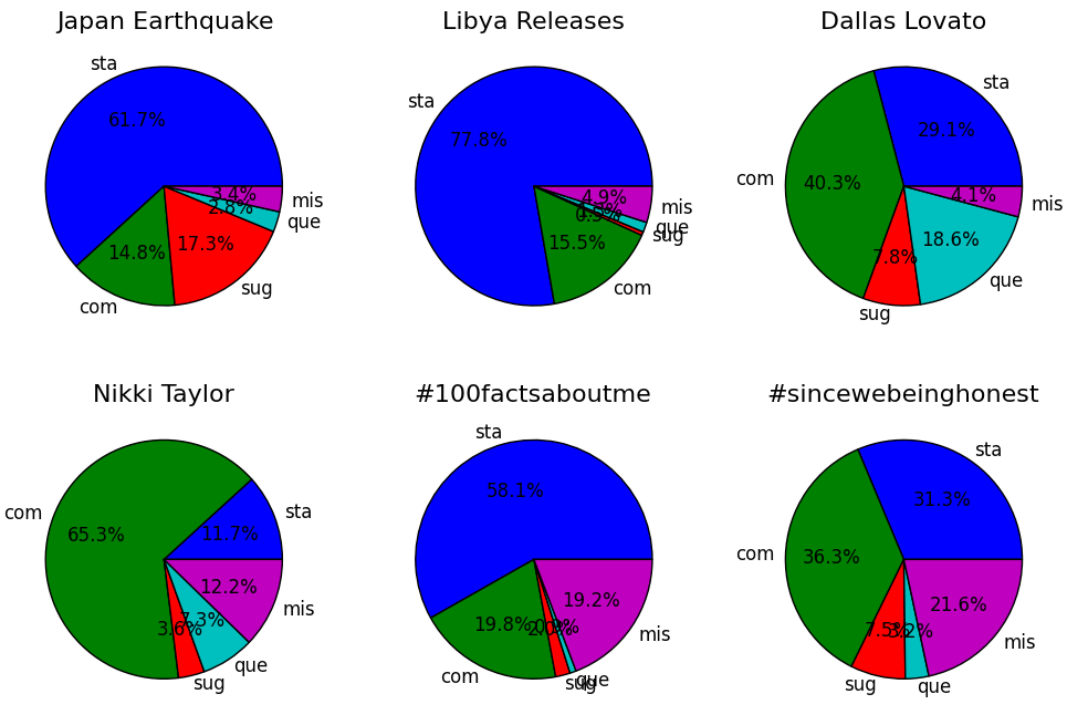


Fig. 2. Speech Act Distributions in the 6 Twitter Topics

The raw Twitter text data were lightly preprocessed and the features were extracted by regular expression patterns. We did two sets of experiments. In the first set, we classified tweets in each topic using different feature sets. The classifier we used is SVM with a linear kernel. Since SVM inherently does binary classification, we solve the multi-class problem by using the one-vs-all paradigm. In the second set, we applied the best feature set from the previous results to three different levels of dataset.

2) Results

For all classification tasks, we report the F1 (the harmonic mean of precision and recall) scores from ten-fold cross validation.

- **Comparison of Feature Sets**

To find out what features are useful for SAR, we experimented with cue words, non-cue features, symbols (symbol-based features), and all our proposed (combined, i.e., cue + non-cue + symbols) features. We also used the commonly adopted bag-of-words (BOW) features for comparison. After removing words that occur only once, we come up with a total of 4421 words as BOW features.

Table IV lists the F1 scores on each speech act type with different feature sets, as weighted averages of the 6 topics according to tweet numbers. The "AVG" is the weighted average according to the number of each speech act type.

TABLE IV
F1 SCORES FOR DIFFERENT FEATURE SETS

Feature set	# Features	Sta	Que	Sug	Com	Mis	AVG
Cue	531	0.788	0.455	0.554	0.623	0.422	0.668
Non-cue	4	0.671	0.088	0.068	0.355	0.074	0.447
Symbols	8	0.681	0.473	0.039	0.412	0.097	0.483
Combined	543	0.798	0.597	0.564	0.670	0.446	0.695
BOW	4421	0.788	0.430	0.533	0.620	0.486	0.673

Among our proposed feature sets, cue words and phrases are the best overall. On individual speech acts, it defeats non-cue words and symbols with the only exception of "questions" because the punctuation *?* is a more reliable indicator of questions than question cue words. Symbol-based features (symbols) outperform non-cue features in almost all columns (with the only exception of "suggestion") and occasionally defeat cue features for reasons explained. Since the non-cue features are meta-features derived from non-cue words bearing the characteristics of cyber English, they are less capable of capturing speech act regularities in Twitter than special symbols. Such evidence also shows that the Twitter text has a distinct style and not all purported "noises" are noisy (e.g., *?!?*).

Without exception, using all our proposed feature sets achieves better performance than using any feature set alone. The combined feature set also defeats the much larger BOW feature set. With a small fixed size, the combined feature set promises good scalability of Twitter SAR. It will be our choice in subsequent experiments.

For individual speech act types, statements and comments are better recognized than questions and suggestions, partly attributable to the difference in training data amount. Unsurprisingly, the recognition of "miscellaneous" is the worst using our features because our proposed features are

aimed to capture the textual characteristics of speech acts, which do not exist in a heterogeneous group made up of different speech act types and non-speech acts. Note that the inferiority in this "speech act" has no adverse effect on our work based on recognized speech acts since no useful information will be derived from it.

- **Comparison of SAR on Different Levels of Dataset**

It is interesting to find out a desirable level to do SAR on – topic-level, category-level, or Twittersphere-level. We expect it to be a higher one because that means we don't have to prepare training data for specific topics or categories, thus simplifying the building of practical systems and saving much annotation labor. Drawing on the previous empirical results, we performed Twitter SAR using the combined feature set on the three levels of dataset, with the results summarized in Table V.

TABLE V
WEIGHTED AVERAGE F1 SCORES ON THREE LEVELS OF DATASETS

Level of Dataset	Sta	Que	Sug	Com	Mis	AVG
Topic	0.798	0.597	0.564	0.670	0.446	0.695
Category	0.673	0.705	0.581	0.629	0.335	0.673
Twittersphere	0.770	0.636	0.577	0.612	0.209	0.639

The average score for all speech act types on the category level or Twittersphere level is not much worse than on the topic-level weighted, degrading only 3% or 8%. The scores on "questions" and "suggestions" are even higher on the category and Twittersphere levels, suggesting that merging data from different topics or categories helps to capture more characteristics of those speech acts. Degradation for "miscellaneous" is attributable to the reasons explained before. But no harm from the "miscellaneous" failure will be inflicted on the work to be presented in the next two sections.

Those evidences enable us to do Twitter SAR on the most general Twittersphere level, without substantial loss in classification performance and with the benefit of using all our annotated data (over 8000 tweets) and obviating the effort to determine the content domain of unseen data.

IV. Speech Act-Guided Key Word/Phrase Extraction

The purpose of doing Twitter SAR is to sort out the tweeted content for extracting summary-worthy information. Among the 5 recognized speech acts, we focus on only 4 "real" types (statement, comment, suggestion, question) and extract key phrases and words from the tweets of major speech act types because they are representative of all communications under the topic. In our experiment, we define "major speech act types" to be those covering at least 20% of all the topic tweets.

The introduction of speech acts facilitates a high-level and well-organized view of the tweets, i.e., whether most of them are about facts, opinions, suggestions, or questions. On this level, we can extract particular language expressions to convey the most salient information in a speech act, which would not be feasible with a more traditional framework working with salient terms, phrases, sentences, or tweets in general.

A. Noise-resistant Phrase Extraction

To extract key words and phrases from the tweets of major speech act types, we first compile a stopword list to filter less informative words. Since general stopword lists such as (Salton 1971) are targeted at standard English, we augment the general stopword list with Netspeak-style acronyms and abbreviations using the free resources mentioned in III[10]. Then we extract key words as frequent nonstop words. Extracting the key phrases is formulated as finding frequent ngram collocations.

Many approaches to collocation finding are based on statistical tests, such as t-test and chi-square test. We use likelihood ratio, a statistical test that gives the ratio of a non-collocation (word independence) likelihood to a collocation (word dependence) likelihood. It has been shown (Dunning 1993) that likelihood ratio does not assume a normal distribution as t-test does and it is more

[10] This Twitter stopword list contains 1760 unique tokens and will be made publicly available.

appropriate for sparse data (e.g., text ngrams) than chi-square.

Regarding an ngram, for two hypotheses H_0 = the occurrences of the *n* words are independent and H_1 = the occurrences of the *n* words are dependent on each other, we use $L(H)$ to represent the likelihood and calculate $\log(L(H_0) / L(H_1))$. Likelihoods are calculated using n-nomial distribution and ngram probabilities are estimated using MLE. For each topic, we extract 50 top bigram phrases, 50 top trigram phrases, and as many "longer phrases" ($n > 3$) as possible with the highest likelihood ratios. There are no more than 10 longer phrases in most cases and their length is typically 4, such as *Appeals Programme Illegal Arrest*.

The collocation-based phrase extraction is resistant to Twitter noise because noisy text by nature is accidental and un-conventionalized. Tweeters produce different kinds of noisy text so that hardly a single noisy phrase appears frequently enough to be extracted by our method. We manually checked 100 randomly sampled key phrases and confirmed that all of them are meaningful and noise-free.

B. POS-based Phrase/Word Patterns

Not all the extracted key words and phrases convey the most relevant information to a speech act. For example, statements are about facts, things, people, etc. and suggestions are about actions, activities, etc. Such information can be approximated by part-of-speech (POS) patterns for both words and phrases. Representative POS-based regular expression patterns are listed in the following, along with illustrative examples.

- The statement-relevant word is a noun, or '/N/' (e.g., *school*), phrase is a noun phrase, such as '/Adj/ /N/' (e.g., *high quality*) and '/Adj/ /N/ /N/' (e.g., *sexual abuse charges*).
- The comment-relevant POS patterns are like the statement-relevant ones. But comment phrases must have at least one opinion word (e.g., *good thing*) judged from SentiWordNet [34] and the

Wilson Lexicon [35].

- The suggestion-relevant word is a verb, or '/V/' (e.g., *hate*), phrase is verb-centered[11], such as '/Adv/ /V/' (e.g., *truly wish*) and '/V/ /N/ /N/' (e.g., *sell health drugs*).
- The question-relevant word is either a verb or a noun, or (/'N'/ | /'V'/) (e.g., *reason*), phrase is either a noun phrase or a verb-centered phrase, such as '/Adj/ /N/ /N/' (e.g., *dirty ass mirror*).

The POS-based extraction is easy to implement and robust in the face of Twitter's noisy text – for which deep NLP such as syntactic or semantic parsing is not appropriate.

C. Phrase/Word Ranking

Among the speech act-relevant words and phrases (ngrams) we only select the most salient ones for a summary. In our work, "salience" is understood as a cumulative effect from an ngram network, i.e., a salient ngram co-occurs with other salient terms in the same tweet, which in turn boosts the salience of other ngrams it co-occurs with.

Let's construct a graph G for the whole tweets of a major speech act type, using all the extracted ngrams (Ng) as vertices. Two vertices Ng_i and Ng_j are linked by an edge if they co-occur in some tweet and the weight of the edge (w_{ij}) is the number of such co-occurrences. Note that G is undirected and we use $NB(Ng_i)$ to denote the neighborhood of Ng_i. Then we define the graph score of Ng_i, $GS(Ng_i)$, as:

$$GS(Ng_i) = \frac{1-d}{|Ng|} + d \times \sum_{Ng_j \in NB(Ng_i)} \frac{GS(Ng_j) \times w_{ij}}{\sum_{Ng_k \in NB(Ng_j)} w_{kj}}$$

The calculation is iterated until convergence. As is the usual practice [37], d is set to be 0.85. This formulation basically follows the TextRank algorithm [20] that can apply to summarization. But their graph vertices are all unigrams from which phrases are later assembled, whereas our Ng includes

[11] It is so called to avoid being confused with the "verb phrase" in a syntactic sense, which is actually a kind of verb-centered phrase.

ngrams of different lengths, which are scored in one process.

Although the extracted phrases are noise-resistant, the same is not true about the extracted words as frequent unigram noises do exist. Moreover, phrases are more informative and less ambiguous than words (compare *school life* with *school* or *life*) and longer phrases are more so. Therefore we count the length *Ni* of Ng_i into its salience score $SS(Ng_i)$, thus rewarding longer ngrams: $SS(Ng_i) = GS(Ng_i) \times N_i$ and rank all the phrases above all the words. Within all the phrases and all the words, rankings are determined by salience scores.

V. Twitter Topic Summarization

For a Twitter topic, the salient words/phrases extracted for its major speech act types as well as the topic itself are the building blocks of a summary. The summary is abstractive in nature as proper words/phrases are to be filled in slots of a template specially designed to accommodate (English) speech acts and speech act verbs. In this section, we first address the missing building block – topic words – which is nontrivial for hashtag topics. Then we provide details of template design and propose a novel summarization algorithm for Twitter topics.

A. Topic Processing

A Twitter topic is itself important information that should be included in the summary because it represents the common ground – sometimes the only common ground – shared by all its tweets. For a **regular topic** in words and phrases like *Space Shuttle*, the inclusion of topic words is straightforward and trivial. For a **hashtag topic** as # plus a concatenation of non-delimitated characters like *#justinbieber*, it is less so. We now describe how to split a hashtag into words.

To begin with, we identify two major types of hashtags, those with mixed-case characters such as *#CyberMonday* and those with all lower-case characters like *#letsbehonest*. The first type resembles

the "upper camel casing" naming convention familiar to programmers, which is easy to detect and split with a simple heuristic. To split the second type and sometimes the result after applying the heuristic to a mixed-case hashtag (e.g., *#PrayforRickRoss*), we rely on the mature statistical-based method successfully applied to other similar tasks such as Chinese word segmentation [38]. To obtain ngram statistics, we count both unigrams and bigrams from all tweets used in our experiments (100 regular topics + 100 hashtag topics, with 5000 tweets in each topic), totaling about 2GB text data and 2.3 million unigrams and bigrams.

After removing the #, we consider every splitting $f = (w_1, w_2, \ldots w_n)$ of a hashtag by scoring it with ngram statistics: $score(f) = (s_{ug}(f) + \lambda s_{bg}(f)) \times lp(f)$ where $s_{ug} = \sum_{i=1}^{n} \log(P(w_i))$ and $s_{bg} = \sum_{i=1}^{n-1} \log(P(w_i w_{i+1}))$. They represent the unigram-based score and bigram-based score of f and λ determines the relative weight of bigrams. The probabilities $P(w_i)$ and $P(w_i w_{i+1})$ are estimated from the corpus using smoothed MLE. We penalize long words by $lp(f)$, which equals 1 if the average word length of f, $wl(f)$, is no more than r; otherwise $lp(f) = wl(f) / r$.

Suppose a hashtag H has m characters, H_k represents the first k characters of H and $Split(H_k)$ the best splitting of H_k. The hashtag splitting algorithm is based on dynamic programming and shown in Fig. 3. Its complexity is $O(m^2)$.

$Split(H_0)$ is empty; $Split(H_1)$ is H's first character itself;
For i = 2 to m
    For j = 0 to i – 1
        Calculate $score(f_j)$ where f_j is formed by $Split(H_j)$ and a "word" as the remaining part of H_i, with H_j removed;
        Choose the highest scoring f_j to be $Split(H_i)$;
Output $Split(H_m)$ i.e., $Split(H)$;

Fig. 3. Splitting Algorithm for Hashtag Topics

We implemented the splitting algorithm on the 100 hashtags from our experimental dataset (λ = 0.01, $\varepsilon = 10^{-10}$, r = 5). The accuracy is 97%. In the only 3 hashtags not correctly split, one is an

acronym hashtag (*#abdc*) that should be treated as a whole, and in the other two, one word is split into two (*lesson* → (*less*, *on*), *justin* → (*just*, *in*)).

B. Template Design

With the topic words and the salient words/phrases for each major speech act type, we can generate an abstractive summary by inserting them into proper slots of speech act-guided templates. In the current work, we aim at short (tweet-long) summaries, which can be conveniently expressed as sentences. So an apt template corresponds to a grammatical sentence, shown in the following.

For "<topic words>**", people** <verb frame> **"**<ngrams>**"**{**,** (**and**) <verb frame> **"**<ngrams>**"**}*.

Fig. 4. Summary Template

In Fig. 4, boldfaced words and punctuations are template constants and the angle brackets (< >) enclose template slots to be filled; (**and**) means the word **and** is optional; { }* means the enclosed part can appear zero or one or more times. The "topic words" are derived from the topic. For a regular topic, they are a direct copy; for a hashtag topic, they are the split result of the hashtag. The "ngrams" are the salient words/phrases extracted for the major speech act types. A "verb frame" is a verb or verb phrase specific to a particular speech act type. For the 4 speech acts types used, we choose typical and short expressions from a lexicon of English speech act verbs [33] listed below.

TABLE VI
VERB FRAMES FOR THE SPEECH ACT TYPES

Speech act type	***Verb frame***
Statement	*state*
Comment	*comment on*
Suggestion	*suggest*
Question	*ask about*

The verb frames are designed to agree with the POS patterns of the succeeding words/phrases to form grammatical constructions such as *state NP*. Problems arise with *ask about* because what follows may not be a verb (phrase) and even so the verb may not be gerundive, which is also a problem for

suggest. POS tagger errors will further break the grammaticality of the sentence. To alleviate such problems, we introduce quotation marks (“ ”) around the succeeding words/phrases so that the sentence generally becomes more readable (compare ... *suggest do your homework* and ... *suggest “do your homework”*).

This is expedient, robust, but not perfect. A better strategy is to lock down individual nouns and verbs and apply proper verb conjugations (+*ing*), a key issue for future exploration.

C. Summarization Algorithm

Each <verb frame> **“**<ngrams>**”** clause in the template represents the salient information about one speech act. We first decide the specific verb frames according to all the major speech act types and order them in the template according to the number of tweets with the speech acts. For example, if a topic has only two major speech act types: “statement” and “comment” with 2000 and 2500 tweets respectively, the template is “For …, … people comment on …, and state …”

The next step is to derive the ngrams needed for the template. We select the ngrams belonging to different speech acts in a round-robin fashion. Starting from the first speech act type as reflected in the order of the verb frames in the template, we select the top-ranking ngrams to fill in template slots. After the last speech act type is processed and if the summary length limit is not reached, we loop back to the first speech act type. The detailed algorithm is presented in Fig. 5.

```
Repeat
    For each speech act in the template order
        Select the top-ranking Ng* from all the ngrams extracted for that speech act;
            If Ng* is a unigram
                Skip to the next speech act unless all longer ngrams (length ≥ 2) for all speech
                acts have been selected;
            If Ng* is not redundant and summary length permits
                Fill a template slot with Ng*;
            Else
                Remove Ng*;
Until summary length is reached;
```

Fig. 5. Ngrams Selection Algorithm

Our algorithm consistently favors longer ngrams so that the generated summary contains informative and less ambiguous phrases. As in multi-document summarization in general, information redundancy should be avoided [39]. We decide whether an ngram is redundant by comparing its words with each of the selected ngrams as well as the topic words. Suppose Ng_0 is selected and Ng_1 is under consideration, we use $W(Ng_0)$ to denote the word set of Ng_0 and decide Ng_1 is redundant if $\frac{|W(Ng_0) \cap W(Ng_1)|}{|W(Ng_0) \cup W(Ng_1)|} \geq \theta$. θ is 0.35 in our experiment.

Note that the template-based approach allows character-level length control. So unlike truncation methods that may leave the last sentence unfinished or last word incomplete, our summarization algorithm guarantees the completeness and readability of the generated summaries.

VI. Summarization Evaluation

In this section, we report how the abstractive summaries generated by our proposed method compare with other automatic summaries and human summaries, as evaluated both automatically and manually.

A. Data Preparation

Using the Twitter search API, we collect trending topic tweets over a one-year period from March 1, 2011 to February 29, 2012. From those topics we construct two datasets: one for regular topics and the other for hashtag topics, each with 100 trending topics covering a variety of categories. Regular topics include news (e.g., *Frankfurt Airport*), entertainment (e.g., *Grammys*), celebrities (e.g., *Jeremy Lin*), technology (e.g., *Android 5.0*), social life (e.g., *Earth Hour*), etc. Hashtag topics include personal life (e.g., *#oomf*), chitchat (e.g., *#idontunderstandwhy*), social life (e.g., *#teaparty*), entertainment (e.g, *#idol*), etc. For each topic we collect up to 5000 distinct tweets (with unique tweet

IDs), with a total of 1 million tweets.

As in SAR evaluation, we require no text cleaning or normalization for the raw tweets. The only NLP tool we need is a POS tagger trained on tweet data [40][12]. The human summaries are collected from two public services: *www.whatthetrend.com* and *tagdef.com*. The former asks users to explain why a topic is trending and the latter, dedicated only to hashtag topics, asks user to define a hashtag topic. The explanations or definitions are required to be short[13] and informative, thus good surrogates for "summaries" in the lack of authentic summaries.

We can find a short explanation or definition of all the 100 regular topics (on *whatthetrend.com* only) and the 100 hashtag topics (on either *whatthetrend.com* or *tagdef.com*), and usually there are multiple versions on one service. Fortunately, both services provide peer check mechanisms to help us choose the best version. *Whatthetrend.com* allows users to verify the posted explanations. *Tagdef.com* employs a voting scheme, allowing users to vote for ("upvotes") or against ("downvotes") a definition. Then we can calculate a score (= number of upvotes – number of downvotes) to indicate the quality of the definition. We choose the summary with the highest score (for *tagdef.com*) or the most recently verified (for *whatthetrend.com*) that fits the time span of the collected tweets. If none of the version is peer-checked, we simply choose the one that best fits the time span. For a regular topic, we can only choose among the versions from the *whatthetrend.com* source. A hashtag topic summary may come from one or two sources. If both sources provide a candidate summary for a hashtag topic and only one is peer-checked, that becomes our human summary. Otherwise we choose the one that best fits the time span.

[12] We thank an anonymous reviewer of an early version of this paper for directing us to this work.

[13] *Whatthetrend.com* limits the length to 140 characters and *tagdef.com* has similar requirement.

B. *Automatic Evaluation*

For comparison, we generate peersummaries of two kinds. The first is by SumBasic, a simple but very robust extractive summarizer for generic documents [41]. The second is by "Hybrid TF-IDF" [23] that ranks tweet sentences by the normalized TF-IDF of their words, a simple system that reportedly defeats MEAD, LexRank, and TextRank for Twitter topic summarization [24]. To ensure fairness, all automatic summaries are no more than a tweet long ($\leq$ 140 characters), as are the human summaries.

For automatic evaluation, we use the popular ROUGE metric [42] to measure the ngram overlap between automatic summaries and human summaries. Popular ROUGE scores used in open tests or competitive events [14] are **ROUGE-1** (unigram overlap), **ROUGE-2** (bigram overlap), and **ROUGE-SU4** (skip bigram overlap, with up to 4 words as the skip distance). Tables VII and VIII report the average ROUGE-1, ROUGE-2, and ROUGE-SU4 F scores for regular topics and hashtag topics respectively. Each score is accompanied by the 99% confidence interval calculated by the ROUGE tool [42]. Statistical significance ($p < 0.01$) under the paired t-test between the peer methods and our method is marked by *.

TABLE VII
ROUGE F SCORES FOR THE REGULAR TOPICS

	ROUGE-1	*ROUGE-2*	*ROUGE-SU4*
Our method	**0.1903** (*0.1642 - 0.2191*)	**0.0588** (*0.0438 - 0.0746*)	**0.0555** (*0.0444 - 0.0661*)
SumBasic	*0.1332 (*0.1114 - 0.1541*)	*0.0440 (*0.0310 - 0.0576*)	*0.0419 (*0.0322 - 0.0527*)
Hybrid TF-IDF	*0.1613 (*0.1353 - 0.1919*)	0.0558 (*0.0386 - 0.0776*)	0.0539 (*0.0399 - 0.0723*)

TABLE VIII
ROUGE F SCORES FOR THE HASHTAG TOPICS

	ROUGE-1	*ROUGE-2*	*ROUGE-SU4*

[14] See DUC (http://www-nlpir.nist.gov/projects/duc/pubs.html) and TAC (http://www.nist.gov/tac/) summarization track guidelines.

Our method	**0.1269** (*0.1039 - 0.1511*)	**0.0357** (*0.0228 - 0.0486*)	**0.0380** (*0.0282 - 0.0482*)
SumBasic	*0.0659 (*0.0457 - 0.0863*)	*0.0074 (*0.0013 - 0.0168*)	*0.0170 (*0.0103 - 0.0249*)
Hybrid TF-IDF	*0.0673 (*0.0473 - 0.0881*)	*0.0134 (*0.0039 - 0.0253*)	*0.0193 (*0.0117 - 0.0286*)

Obviously, our proposed method leads in all ROUGE measures on both types of topics. Consistent with the results reported in previous work [26], regular topic summaries are much better than hashtag topic summaries. Also note that it is on the hashtag topic summaries that our proposed method more markedly excels, significantly outperforming SumBasic and Hybrid TF-IDF.

C. Manual Evaluation

TABLE IX
HUMAN AND AUTOMATIC SUMMARIES FOR *#agoodboyfriend*

Human	*People are tweeting the qualities that make a good boyfriend and the things a good boyfriend does.*
Our method	*For "a good boyfriend", people state "Team Minaj, DAMN Derrick Rose, Yuri Gagarin" and comment on "love joy, silent cries, good girlfriend".*
SumBasic	*#agoodboyfriend is #agoodboyfriend whether he's around u or not.. "#AGoodBoyfriend" is really a TT ? #agoodboyfriend is not looking for #ago*
Hybrid TF-IDF	*RT @DamnItsTrue: GREAT LIFE = Good Friends Good Food Good Song #agoodboyfriend #DamnItsTrue @DamnItsTrue: GREAT LIFE = Good Friends +*

A closer inspection of the summaries reveals that the abstractive summaries guided by speech acts more often capture key words or phrases in human summaries than the extractive summaries, which are vulnerable to spam, redundancy, and other noisiness as shown in Table IX, which lists the human and automatic summaries for the hashtag topic *#agoodboyfriend*.

In addition to key word overlapping (“people”, “good”, “boyfriend”), our abstractive summary are structurally similar to the human summary (“people are tweeting …” vs. “people state …”) and both are expressed in complete sentences. On the contrary, the two extractive summaries are only

combinations of tweets or tweet fragments. In addition, the contents of the human summary and the automatic summaries are of different kinds, so that the ngram overlap is generally low for all the automatic summaries. Unlike the very general tendencies in human summaries ("qualities", "things") and the very specific but trivial messages in extractive summaries (from tweets), our abstractive summaries provide general ("good") as well as specific information ("Derrick Rose", "Yuri Gagarin" as examples of "good boyfriend"). Whether this makes good summaries cannot be directly evaluated by ROUGE.

For such concerns, we also do manual evaluations on three criteria – **explanatoriness**, **informativeness**, and **readability**. Informativeness and readability are generally accepted yardsticks for a summary's content and form. Explanatoriness is also introduced as our human summaries are primarily explaining what #XXX is (*tagdef.com*) or why #XXX is trending (*whatthetrend.com*).

We trained two human judges to score the summaries according to the criteria described above on a scale of 5 points. The higher the score, the more explanatory / informative / readable a summary is. Each judge is required to score all the human and automatic summaries, totaling $100 \times 4 \times 2 = 800$ summaries. For each topic, they are presented the summaries in a randomly scrambled order so that no pattern can be detected. For each scoring category, Cohen's Kappa ranges between 0.5 and 0.7, indicating good inter-judge agreement. Tables X and XI sum up the results on the regular and hashtag topics by averaging the human scores over 100 topics each, with statistical significance of the summaries generated by our method against all the other summaries indicated by * ($p < 0.001$) on a paired two-tailed t-test.

TABLE X
AVERAGE HUMAN SCORES FOR THE REGULAR TOPICS

	Explanatoriness	*Informativeness*	*Readability*
Human	4.07	3.84	*4.71

Our method	**3.98**	**3.78**	**4.01**
SumBasic	*2.35	*1.88	*2.60
Hybrid TF-IDF	*2.41	*1.95	*2.25

TABLE XI
AVERAGE HUMAN SCORES FOR THE HASHTAG TOPICS

	Explanatoriness	*Informativeness*	*Readability*
Human	3.61	3.26	*4.63
Our method	**3.44**	**3.19**	**3.61**
SumBasic	*2.23	*1.94	*2.55
Hybrid TF-IDF	*2.55	*2.17	*2.65

The statistics show that the summaries generated with our method are comparable to human writings in terms of explanatoriness and informativeness. On these criteria our method significantly outperforms SumBasic and Hybrid TF-IDF with a large margin. The same is also true for readability, showing the superiority of abstractive summarization. But our summaries are also significantly less readable than human writings, mainly because of the lack of coherence between the extracted key words and phrases. For example, there is no link between "love joy, silent cries, good girlfriend" in the sample summary (Table IX). The regular topic summary scores are generally higher than the corresponding hashtag topic scores, which is consistent with the ROUGE results shown in Tables VII and VIII and lends further credence to the robustness of our method.

VII. CONCLUSIONS AND FUTURE WORK

In this paper, we have taken a new initiative for Twitter topic summarization – speech act-guided summarization. To automatically recognize speech acts in tweets, we treat SAR as a multi-class classification problem and propose a set of word-based and symbol-based features that can be easily harvested from raw data or free resources. With the recognized speech acts, we proceed to extract key words and phrases from tweets to compose abstractive summaries, with the aid of the noise-resistant phrase extraction and POS pattern filtering. The extracted key terms are then ranked and inserted into

special summary templates designed for speech acts, using a round-robin algorithm. Designed to accommodate the numerous, short, dissimilar, and noisy nature of the summarization object – tweets, our approach makes a solid contribution to the summarization community.

Evaluated automatically on two types of Twitter trending topics, regular and hashtag, our proposed summarization method outperforms two representative extractive summarizers. Manual evaluation results show that our abstractive summaries are significantly more explanatory, informative, and readable than the two kinds of extractive summaries and comparable to the human counterparts in terms of both explanatoriness and informativeness.

In the future, we are going to improve Twitter SAR by experimenting with different classifiers, especially the inherent multi-class types such as Naïve Bayes and Decision Tree. As human labeling is expensive and time-consuming, research in a semi-supervised approach is also underway. The summarization framework can also be improved, especially in summary readability. A promising venue is to incorporate the context of key words and phrases during their extraction and count contextual similarity or co-occurrence frequencies into the ranking and template-filling of the extracted terms.

ACKNOWLEDGEMENT

We are very grateful to the help by Jian Xu and Oscar Lai. The work described in this paper was supported by the grant GRF PolyU 5230/08E.

REFERENCES

[1] J. AUSTIN, *How to Do Things with Words*. Oxford: Oxford University Press. 1962.

[2] J. SEARLE, “Indirect Speech Acts”. In P. Cole and J. Morgan (Eds.), Syntax and semantics, vol. iii: Speech acts (pp. 59–82). New York: Academic Press. 1975.

[3] J. SEARLE and D. VANDERVEKEN, *Foundations of Illocutionary Logic*. Cambridge: Cambridge University Press. 1985.

[4] H. BUNT, "Context and Dialogue Control". *Think*, 3: pp. 19–31. 1994.

[5] A. STOLCKE, K. RIES, N. COCCARO, E. SHRIBERG, R BATES, D. JURAFSKY, P. TAYLOR, R. MARTIN, C. VAN ESS-DYKEMA, and M. METEER, "Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech". COMPUT LINGUIST. vol. 26, no. 3, pp. 339–373, 2000.

[6] E. SHRIBERG, R. BATES, P. TAYLOR, A. STOLCKE, D. JURAFSKY, K. RIES, N. COCCARO, R. MARTIN, M. METEER, and C. VAN ESS-DYYKEMA, "Can Prosody Aid the Automatic Classification of Dialog Acts in Conversational Speech?" LANGUAGE AND SPEECH (SPECIAL ISSUE ON PROSODY AND CONVERSATION), vol. 41, no. 3–4, pp. 439–487. 1998.

[7] D. JURAFSKY, E. SHRIBERG, and D. BIASCA, "Switchboard SWBD-DAMSL Labeling Project Coder's Manual, Draft 13". Technical report, University of Colorado Institute of Cognitive Science. 1997.

[8] R. DHILLON, S. BHAGAT, H. CARVEY, and E. SHRIBERG, "Meeting Recorder Project: Dialog Act Labeling Guide". Technical report, International Computer Science Institute. 2004.

[9] W. COHEN, V. CARVALHO, and T. MITCHELL, "Learning to Classify Email into 'Speech Acts'". In Proc. EMNLP-04, pp. 309–316. 2004.

[10] D. FENG, E. SHAW, J. KIM, and E. H. HOVY. "Learning to Detect Conversation Focus of Threaded Discussions". In Proc. HLT-NAACL-06, pp. 208–215. 2006.

[11] R. ZHANG, D. GAO, and W. LI, "What Are Tweeters Doing: Recognizing Speech Acts in Twitter". In AAAI-11 Workshop on Analyzing Microtext. 2011.

[12] A. RITTER, C. CHERRY, and B. DOLAN, "Unsupervised Modeling of Twitter Conversations". In Proc. HLT-NAACL-10, pp. 172–180. 2010.

[13] M. JEONG, C-Y. LIN, and G. LEE, "Semi-supervised Speech Act Recognition in Emails and Forums". In Proc. EMNLP-09, pp. 1250–1259. 2009.

[14] D. CRYSTAL, *Language and the Internet. 2nd edition*. Cambridge: Cambridge University Press. 2006.

[15] D. CRYSTAL, *Internet linguistics*. London: Routledge. 2011.

[16] M. KAUFMANN and J. KALITA, "Syntactic Normalization of Twitter Messages". In Proc. ICON-2010: 8th International Conference on Natural Language Processing. 2010.

[17] B. HAN and T. BALDWIN, "Lexical Normalisation of Short Text Messages: Makn Sens a #twitter". In Proc. ACL-11, pp. 368–378. 2011.

[18] K. MCKEOWN and D. R. RADEV, "Generating Summaries of Multiple News Articles". In Proc. 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 74–82. 1995.

[19] D. RADEV, H. JING, M. STYŚ, and D. TAM, "Centroid-Based Summarization of Multiple Documents". INFORMATION PROCESSING AND MANAGEMENT, vol. 40, pp. 919–938. 2004.

[20] R. MIHALCEA and P. TARAU, "TextRank: Bringing Order into Texts". In Proc. EMNLP-04, pp. 404–411, 2004.

[21] X. WAN, and J. YANG, "Multi-Document Summarization Using Cluster-Based Link Analysis". In Proc. SIGIR-08, pp. 299–306. 2008.

[22] B. SHARIFI, M-A. HUTTON, and J. KALITA, "Summarizing Microblogs Automatically". In Proc. HLT/NAACL-10. 2010.

[23] B. SHARIFI, M-A. HUTTON, and J. KALITA, "Experiments in Microblog Summarization". In Proc. IEEE Second International Conference on Social Computing. 2010.

[24] D. INOUYE, "Multiple Post Microblog Summarization". REU Research Final Report. 2010.

[25] G. ERKAN and D. RADEV, "LexRank: Graph-based Centrality as Salience in Text Summarization". JOURNAL OF ARTIFICIAL INTELLIGENCE RESEARCH, vol. 22, pp. 457–479. 2004.

[26] F. LIU, Y. LIU, and F. WENG, "Why is 'SXSW' Trending? Exploring Multiple Text Sources for Twitter Topic Summarization". In Proc. of the ACL Workshop on Language in Social Media (LSM 2011), pp. 66–75. 2011.

[27] E. REITER and R. DALE, *Building Natural Language Generation Systems*. Cambridge: Cambridge University Press. 2000.

[28] K. VAN DEEMTER and J. ODIJK, "Context Modelling and the Generation of Spoken Discourse". SPEECH COMMUNICATION, vol. 21, no. 1/2, pp. 101–121. 1997.

[29] S. W. MCROY, S., CHANNARUKUL, and S. S. ALI, "An Augmented Template-based Approach to Text Realization". NATURAL LANGUAGE ENGINEERING, vol. 9, no. 4, pp. 381–420. 2003.

[30] H. SAGGION and G. LAPALME, "Generating Indicative-Informative Summaries with SumUM". COMPUTATIONAL LINGUISTICS, vol. 28, no. 4 pp. 497–526. 2002.

[31] G. CARENINI, and J. C. K. CHEUNG, "Extractive vs. NLG-based Abstractive Summarization of Evaluative Text: The Effect of Corpus Controversiality". In Proc. ACL-08, pp. 33–40. 2008.

[32] M. KUMAR, D. DAS, S. AGARWAL, and A. RUDNICKY, "Non-textual Event Summarization by Applying Machine Learning to Template-based Language Generation". In Proc. of the 2009 workshop on language generation and summarisation (UCNLG+Sum 2009), pp. 67–71. 2009.

[33] A. WIERZBICKA, *English Speech Act Verbs: A Semantic Dictionary*. Orlando: Academic Press. 1987.

[34] S. BACCIANELLA, A. ESULI, and F. SEBASTIANI, "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining". *Proc. the Seventh conference on International Language Resources and Evaluation.* 2010.

[35] T. WILSON, and J. WIEBE, "Identifying Opinionated Sentences". In *Proc. NAACL-03*, pp. 33–34, 2003.

[36] X. ZHAO and J. JIANG, "An Empirical Comparison of Topics in Twitter and Traditional Media". *Technical Report*, Singapore Management University School of Information Systems. 2011.

[37] S. BRIN and L.PAGE, "The Anatomy of a Large-scale Hypertextual Web Search Engine," In *Proc. 7th WWW conference*, pp.107–117, 1998.

[38] Z. WU and G. TSENG, "Chinese Text Segmentation for Text Retrieval Achievements and Problems". *JASIS*, vol. 44, no. 9, pp. 532–542. 1993.

[39] J. CARBONELL and J. GOLDSTEIN, "The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries". In *SIGIR-98*, pp. 335–336. 1998.

[40] K. GIMPEL, N. SCHNEIDER, B. O'CONNOR, D. DAS, D. MILLS, J. EISENSTEIN, M. HEILMAN, D. YOGATAMA, J. FLANIGAN, and N. A. SMITH, "Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments". In *Proc. ACL-11*, 2011.

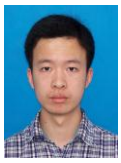
[41] A. NENKOVA and L.VANDERWENDE, "The Impact of Frequency on Summarization". *Technical Report MSR-TR-2005-101*, Microsoft Research, Redmond, WA. 2005.

[42] C-Y. LIN, "ROUGE: A Package for Automatic Evaluation of Summaries". In *ACL 2004* Workshop on Text Summarization Branches, pp. 74–81, 2004.

Renxian Zhang is a PhD candidate in the Department of Computing, the Hong Kong Polytechnic University, Hong Kong. His current research interests include natural language processing, text mining, and document summarization.

Wenjie Li is currently an associate professor in department of computing, the Hong Kong Polytechnic University, Hong Kong. She received her PhD degree from department of systems engineering and engineering management in the Chinese University of Hong Kong, Hong Kong, in 1997. Her main research topics include natural language processing, information extraction and document summarization.

Dehong Gao is a PhD candidate in the Department of Computing, the Hong Kong Polytechnic University, Hong Kong. His current research interests include opinion mining and social network study.



You Ouyang is currently an algorithm engineer in Beijing Miaozhen Systems, China. He received a Ph.D degree of computer science from the Hong Kong Polytechnic University in 2011. His main research interests include statistical natural language processing and text mining.