Question Identification on Twitter

Baichuan Li[1]*, Xiance Si[2], Michael R. Lyu[1], Irwin King[13]†, and Edward Y. Chang[2]
[1]The Chinese University of Hong Kong, Shatin, N.T., Hong Kong
[2]Google Research, Beijing 10084, China
[3]AT&T Labs Research, San Francisco, CA, USA
bcli@cse.cuhk.edu.hk, sxc@google.com, lyu@cse.cuhk.edu.hk
{king@cse.cuhk.edu.hk, irwin@research.att.com}, edchang@google.com

ABSTRACT

In this paper, we investigate the novel problem of automatic question identification in the microblog environment. It contains two steps: detecting tweets that contain questions (we call them "interrogative tweets") and extracting the tweets which really seek information or ask for help (so called "qweets") from interrogative tweets. To detect interrogative tweets, both traditional rule-based approach and state-of-the-art learning-based method are employed. To extract qweets, context features like short urls and Tweet-specific features like Retweets are elaborately selected for classification. We conduct an empirical study with sampled one hour's English tweets and report our experimental results for question identification on Twitter.

Categories and Subject Descriptors

H.3.5 [**Information Systems**]: Online Information Services—*Web-based services*

General Terms

Experimentation, Performance

Keywords

Question Identification, Microblogs, Twitter

1. INTRODUCTION

Twitter[1], as the first and one of the most popular microblog services, has become a platform of question asking. Morris et al. [3] reported that more than 10% of Twitter users once asked questions on Twitter. Furthermore, Efron and Winget [2] found that 13% of their randomly sampled 2-million tweets corpus were questions. According to the blog of Twitter, on average 1,620 tweets were posted every second in March 2011[2], which means each second 210 questions appeared on Twitter.

However, not all of them need to be answered. According to the taxonomy in [2], many questions are not requesting information but providing information, like suggesting a Q&A pair or expressing an opinion. Different from the meanings of questions in [2] and [3], we restrict the scope of questions to be those tweets which require some information or help and thus need to be answered. In the following, we use the term "qweets" to represent these tweets. Additionally, we call the tweets which contain question sentences as "interrogative tweets". It is worth noting that "qweets" are "interrogative tweets" but "interrogative tweets" are not necessary "qweets".

To our best knowledge, few attempts have been made on identifying qweets in vast tweets. Thus, this paper asks such research question: Can we automatically identify qweets from tweets? We know that microblog provides an instant message publishing service, so identifying these questions automatically will help question askers get answers efficiently through approaches such as question routing and automatic question answering system [5]. Identifying qweets also lays a foundation for question analysis (such as classification and clustering) in the microblog environment.

We argue that qweet identification is different from traditional question finding [1, 6]. On the one hand, a tweet contains less than 140 characters, and thus less information (features) is provided; On the other hand, a tweet usually includes some special features, such as Retweets (repost other users' tweets, with "RT @username" as the marker), "@username" (mention or reply to some user), hashtags (terms starting with the charachter "#", usually denote topics) and short urls, which may contribute to qweet identification.

In this paper, we cast the qweet identification problem into a two-phase cascade process: In the first step, we detect the tweets which contain questions (interrogative tweets). We adopt both traditional methods like question marks, 5W1H words, and state-of-the-art approach [1, 6] which utilizes sequential question patterns to detect interrogative tweets. In the second step, we extract qweets by splitting each interrogative tweet into question part and context part and employing four kinds of features (question features, context features, question-context features and tweet-specific features) to build a binary classifier. As an empirical study,

*This work was done when the first author was on internship at Google.

†Irwin King is currently on leave from CUHK to be with AT&T Labs.

[1] http://www.twitter.com/

[2] http://blog.twitter.com/2011/03/numbers.html.

CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK.
Copyright 2011 ACM 978-1-4503-0717-8/11/10 ...$10.00.

we experiment with a data set sampled from one-hour tweet stream and the experimental results demonstrate that: 1) 11% of tweets contain questions and 6% of tweets are qweets; 2) Simple rule-based approach produces satisfactory results in detecting interrogative tweets; 3) Context features and Tweet-specific features are very useful for extracting qweets from interrogative tweets.

The paper proceeds as follows. Section 2 presents the related work. Section 3 details the work of qweet identification. Experiments are described in Section 4 and a conclusion is given in Section 5.

2. RELATED WORK

Detecting Questions in User Generated Content (UGC). Interrogative tweets detection is closely related to question (subquestion) finding in UGC such as forums, community question answering (CQA) portals and microblogs. There are two divisions for question finding in UGC: rule-based approach [2] and learning-based approach [1, 6]. Rule-based approach usually designs several rules from heuristics or observations to check whether a thread or tweet is question or not, while learning-based approach constructs a binary classifier with lexical and/or syntactic features.

Analyzing Questions in Tweets. The study of questions on Twitter is in its infancy and previous work mainly focused on analyzing the types and motivations of questions people asked. Efron and Winget [2] built a taxonomy of questions on Twitter and found that people asked questions on Twitter to both seek information (facts, opinions, etc.) and recommend information (external resources, invitations, QA pairs,etc.). Morris et al. [3] conducted a survey study to reveal the types, and motivations of questions people ask and answer on Twitter and Facebook.

3. QWEET IDENTIFICATION

As mentioned, we break qweet identification into the following two sub-problems: (1) Detecting tweets which contain questions (interrogative tweet detection); (2) Extracting qweets from the previous result (qweet extraction). Detail will be given one by one.

3.1 Interrogative Tweet Detection

Previous work showed that learning-based approach outperformed rule-based approach in question detection in forums [1] and CQA portals [6], so we attempt both rule-based approach and learning-based approach to investigate whether the latter one still outperforms the former in the microblog environment.

3.1.1 Rule-based Approach

We apply question marks, 5W1H words, refined 5W1H and the rules introduced in [2] to detect interrogative tweets, detail will be presented in Section 4.2.1.

3.1.2 Learning-based Approach

This approach utilizes sequential question patterns to train a binary classifier, which needs manually labeling data. To avoid time-costing labeling, we utilize corpus from CQA portals to get mountains of "well labeled" sequential question patterns based on the following two considerations:

- Both tweets and questions in CQA portals are generated by users;
- The length of one question title is usually as short as that of a tweet.

We apply the Prefixspan algorithm proposed by Pei et al. [4] to mine frequent question patterns. This algorithm explores prefix projection to speed up sequential pattern mining. To get high quality question patterns, we also set extra constrains to the original algorithm empirically: The minimum support and minimum confidence are set to be 0.5% and 0.7; The minimum and maximum length of patterns are set to be 2 and 5; The maximum length between two neighbor terms in any pattern is set to be 3. In order to calculate the confidence, we choose the corresponding best answers as negative examples. Furthermore, to avoid long question titles and unbalanced Q&A pair, we limit the length of each question title to be within 5 and 20 words and the maximum length ratio between question title and answer to be 1:5. One-class SVM is employed afterwards to predict whether a tweet contains question(s) or not.

3.2 Qweet Extraction

We present qweet extraction in this section. Firstly, we need to understand which kinds of interrogative tweets are not qweets. From the observations, we construct a taxonomy of interrogative tweets (Section 3.2.1). Afterwards, we extract four types of features according to the characteristics of interrogative tweets and convert qweet extraction to a binary classification problem (Section 3.2.2).

3.2.1 Types of Interrogative Tweets

We classified interrogative tweets to the following 6 types. Only those tweets belonging to the last type are qweets.

1. Advertisement. This kind of tweets ask questions to the reader and deliver advertisements in the following. E.g.,

- *Incorporating your business this year? Call us today for a free consultation with one of our attorneys. 855-529-8753. http://buz.tw/FjJCV*

2. Article or News Title on the Web. These tweets post article names or news titles together with the links to the webpage. E.g.,

- *New post: Pregnancy Miracle - A Miracle or a Scam? http://articlescontentonline.com/pregnancy-miracle-a-miracle-or-a-scam/*

3. Question with Answer. These tweets contain questions followed by their answers. E.g.,

- *I even tried staying away from my using my Internet for a couple hours. The result? Insanity!*

4. Question as Quotation. These tweets contain questions in quoted sentences as references to what other people said. E.g.,

- *I think Brian's been drinking in there because I'm hearing him complain about girls, and then he goes "Wright, are you sure you're not gay?"*

5. Rhetorical Question. This kind of tweets include rhetorical questions, which seem to be questions but without the expectation of any answer. In another words, these tweets encourage readers to think about the obvious answers. E.g.,

- *You ruined my life and I'm supposed to like you?*

6. Qweet. These kinds of tweets ask for some information or help. E.g.,

- *What's your favorite Harry Potter scene?*

Another special type is "Question on The Web" which the

Table 1: Features extracted for qweet extraction

Feature	Description
Question features (Q)	
Quoted question	Whether the question sentence is quoted from other sources
Strong feeling	Whether the question sentence contains strong feeling such as "???" and "?!"
Context features (C)	
URL	Whether the context contains any url
Phone number or Email	Whether the context contains any phone number or email
Strong feeling	Whether there is any strong feeling such as "!" follows the question sentence
Declarative sentence after question sentence	Whether there is any declarative sentence follows the question sentence
Word features	Unigram words appear in the contexts of tweets
Question-Context features (QC)	
Self ask self answer	Whether the tweet contains obvious self ask self answer pattern. E.g., Q:...A:...
Question-url sameness	Whether the question sentence is the same as the webpage's title linked through the url.
Tweet-specific features (T)	
@username	Whether the tweet mentions other user's name.
Retweet	Whether the tweet is a Retweet.
Hashtag	Whether the tweet contains any hashtag.

Tweet author posts a question asked by someone on the web, e.g., CQA portals, forums, etc. The following is an example:

- *Questions about panda update. When will the effect end? http://goo.gl/fb/iiRjn*

However, even for human it is hard to infer the author's intention. Therefore, we classify it to *Type 2* for convenience.

3.2.2 Feature Extraction

Different from traditional short texts like forum posts and CQA questions, tweets own some special characteristics, such as "@username", Retweets, and hashtags. Apart from that, plenty of tweets contain (shortened) links which provide extra information to determine these tweets' types. We also note that an interrogative tweet usually can be split into two parts: question part and context part. Although it is not easy to judge whether a tweet is qweet from the question itself, the context helps to distinguish qweets from non-qweets.

Based on the above findings, we extract four kinds of features from each tweet, which are reported in Table 1. Noting that word overlap similarity is employed to calculate the value of question-url sameness. To be specific, if the overlap similarity between any question sentence and the title of url is above a certain threshold, the value is 1; otherwise we treat them as different content and set the value to 0. In our experiments, we set the threshold to 0.8.

Using these features, we leverage one Random Forrest classifier to predict whether a interrogative tweet is a qweet or not.

4. EXPERIMENTS

4.1 Dataset

We sampled one hour's tweets from 11:00am to 12:00am on April 18, 2011 using Twitter API[3], getting 2,045 English tweets in total. Two raters were asked to label whether each tweet was interrogative tweet. They worked individually and the labels were finalized through discussions between them.

Table 2 reports the statistic of the data, from which we estimate that on Twitter there are about 11% of tweets containing questions (similar to 13% reported in [2]) and 6% of tweets having information needs. This observation confirms the statement that Twitter is no longer a pure social media which let user publish "What's happening"s to provide information but also becomes an online questioning platform where users post "Does anyone know...?"s to seek information.

To extract sequential patterns, we collected over 850,000 question titles and the corresponding best answers from Yahoo! Answers[4] and WikiAnswers[5].

Table 2: Summary of data set

Number of tweets	2,045
Number of interrogative tweets	227
Number of qweets	127
Number of tweets containing urls	442
Number of tweets containing hashtags	459
Number of Retweets	240
Number of tweets containing @	447
Number of tweets containing ?	222
Number of tweets containing 5W1H	293
Average number of words per tweet	12.48

4.2 Experimental Results

4.2.1 Interrogative Tweet Detection

Table 3 presents the accuracy of various methods on detecting interrogative tweets. We find that simple question patterns give satisfactory performance. Specifically, using question mark alone gets the precision of nearly 0.97 but captures less than 85% of all interrogative tweets. Adding 5W1H question words remedies the drawback of low recall, but substantially decreases precision. To reduce the number of false positives, we set up the following two heuristics for 5W1H:

1. They must appear at the beginning of one sentence.
2. Auxiliary words are added to the original words. For example, we change "what" to "what is" and "what are".

The results demonstrate that the above rules boost the performance significantly. The first heuristic improves precision of 5W1H by 60.5% although reduces 5.9% of recall. For F_1, it improves 28.4% comparing with original 5W1H. The second heuristic gives similar results. When applying the above two together, it makes the precision higher than 0.95 and recall higher than 0.9, which increases precision and F_1

[3] http://dev.twitter.com.

[4] http://answers.yahoo.com.

[5] http://wiki.answers.com.

by 7.2% and 3.0% comparing to using question mark with slight decrease on recall (1.5%). We also try the artificial rules designed in [2], which keep high precision but do not increase recall too much.

Learning-based approach, however, does not improve the accuracy of detected questions as expected. Patterns with low confidence get low precision while patterns with high confidence fail to increase recall. We conjecture that for one reason, most of question patterns are "overlapped" with 5W1H, such as "how to" and "where can I". For the other, sequential question patterns in CQA may not well capture the questions on Twitter as many patterns fail to match any tweet. In our future study, we plan to label large number of real tweets to further explore this problem.

Table 3: Accuracies of interrogative tweet detection for various methods (QM: question mark; best results are in bold)

Method	Precision	Recall	F_1
QM	**0.969**	0.846	0.903
QM and 5W1H	0.547	**0.973**	0.700
QM and refined 5W1H (H1)	0.878	0.916	0.899
QM and refined 5W1H (H2)	0.875	0.925	0.899
QM and refined 5W1H (H1 and H2)	0.954	0.907	**0.930**
Rules in [2]	0.960	0.855	0.904
Question Patterns (Confidence $\geq$ 0.7)	0.576	0.899	0.702
Question Patterns (Confidence $\geq$ 0.8)	0.715	0.872	0.786
Question Patterns (Confidence $\geq$ 0.9)	0.857	0.846	0.851

Table 4: Effects of feature sets on qweet extraction

Feature Set	Precision	Recall	Accuracy
Q	0.576	**0.984**	0.577
Q+QC+C	0.704	0.937	0.739
Q+QC+non-word C	0.714	0.906	0.730
Q+QC+non-word C+T	0.764	0.866	0.770
Q+QC+non-word C+Retweet	**0.766**	0.874	**0.775**
Q+QC+non-word C+@username	0.728	0.929	0.761
Q+QC+non-word C+Hashtag	0.702	0.890	0.721

4.2.2 Qweet Extraction

In our experiments, we employed a Random Forrest classifier[6] with 1000 trees to extract qweets and adopted 10-fold cross validation. Table 4 presents the results of qweet extraction using different sets of features. We find that: 1) **Context features are of great importance in distinguishing qweets from non-qweets.** Comparing the classifier's performance with feature sets "Q" and "Q+QC+C", it is obvious that context features boost the performance greatly. To be specific, precision is improved by 22% with a cost of slight decrease in recall (We consider precision as more important than recall since high recall is easier to reach if the classifier predicts all interrogative tweets as qweets). However, unigrams in context seem useless since "Q+QC+non-word C" gives even higher precision than "Q+QC+C". We conjecture that although some words like "free", "click", etc. are useful in judging advertisement and online articles, much more irrelevant words are introduced when using word features. 2) **Tweet-specific features also help in qweet identification.** Comparing the results using features sets "Q+QC+non-word C" and "Q+QC+non-word C+T", we find that Tweet-specific features improve the precision of qweet extraction. In addition, without Tweet-specific features, the classifier's recall increases, which means it tends to predict a interrogative tweet as qweet when such information is missing. From the last three rows of Table 4 we can further understand that Retweet is the most significant feature, followed by "@username", while Hashtag is not so helpful. The possible explanation is that most Retweets are not qweets as they just rebroadcast other users' tweets, which are usually providing rather than asking for information. On the contrary, "@username" always denotes a interrogative tweet being a qweet as it implies a dialog between two users. As our data contain no obvious qweet signals in Hashtags (like "#lazyweb") and Hashtags are usually used to represent the topics, they are not so influential. 3) **Qweet extraction is a non-trivial problem.** At present, the performance is still not so satisfactory. Through examining the failed examples, we find that tweets containing rhetorical questions and complicated self-ask-self-answer sentences are always being misclassified. E.g., "If I'm that mean/nasty why do u continue to talk to me" and "See what you did? Poor 4 minute. They...". Potentially this problem may be solved through utilizing syntax features to capture the structure of tweets and employing more training data.

5. CONCLUSION AND FUTURE WORK

This paper explored the problem of automatic question identification in the microblog environment. To identify qweets, we designed a cascade process which first detected interrogative tweets and then extracted qweets. For interrogative tweet detection, both rule-based approach and learning-based approach were employed and experimental results showed that on Twitter rule-based approach performed well while learning-based approach did not improve the performance significantly. For qweets extraction, we leveraged the special characteristics of Tweets like Retweet and @username, and context of questions to build a binary classifier. Experimental results demonstrated that such features were of great importance in extracting qweets from interrogative tweets. Looking forward, we plan to enlarge our dataset and utilize syntax features in qweet extraction.

ACKNOWLEDGEMENT

This work is supported by two grants from the Research Grants Council of the Hong Kong SAR, China (Project No. CUHK 413210 and Project No. CUHK 415410) and a grant from Google Focused Grant Project "Mobile 2014".

6. REFERENCES

[1] G. Cong, L. Wang, C.-Y. Lin, Y.-I. Song, and Y. Sun. Finding question-answer pairs from online forums. In *Proc. of SIGIR '08*, 2008.

[2] M. Efron and M. Winget. Questions are content: a taxonomy of questions in a microblogging environment. In *Proc. of ASIST '10*, 2010.

[3] M. R. Morris, J. Teevan, and K. Panovich. What do people ask their social networks, and why?: a survey study of status message Q&A behavior. In *CHI '10*, 2010.

[4] J. Pei, J. Han, B. Mortazavi-asl, H. Pinto, Q. Chen, U. Dayal, and M. chun Hsu. Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth. In *Proc. of ICDE '01*, 2001.

[5] X. Si, E. Y. Chang, Z. Gyöngyi, and M. Sun. Confucius and its intelligent disciples: integrating social with search. *Proc. VLDB Endow.*, 3:1505–1516, 2010.

[6] K. Wang and T.-S. Chua. Exploiting salient patterns for question detection and question retrieval in community-based question answering. In *COLING '10*, 2010.

[6] We attempted Random Forrest, SVM, J48, and Logistic Regression in our experiments and Random Forrest performed the best among them.