Why is "SXSW" trending? Exploring Multiple Text Sources for Twitter Topic Summarization

Fei Liu[1] **Yang Liu**[1] **Fuliang Weng**[2]
[1]Computer Science Department, The University of Texas at Dallas
[2]Research and Technology Center, Robert Bosch LLC
{feiliu, yangl}@hlt.utdallas.edu[1]
fuliang.weng@us.bosch.com[2]

Abstract

User-contributed content is creating a surge on the Internet. A list of "buzzing topics" can effectively monitor the surge and lead people to their topics of interest. Yet a topic phrase alone, such as "SXSW", can rarely present the information clearly. In this paper, we propose to explore a variety of text sources for summarizing the Twitter topics, including the tweets, normalized tweets via a dedicated tweet normalization system, web contents linked from the tweets, as well as integration of different text sources. We employ the concept-based optimization framework for topic summarization, and conduct both automatic and human evaluation regarding the summary quality. Performance differences are observed for different input sources and types of topics. We also provide a comprehensive analysis regarding the task challenges.

1 Introduction

User contributed content has become a major source of information in the Web 2.0 era. People follow their topics of interest, share their experience or opinions on a variety of interactive platforms, including forums, blogs, microblogs, social networking sites, etc. To keep track of the trends online and suggest topics of interest to the general public, many leading websites provide a "buzzing" service by publishing the current most popular topics on their entrance page and update them regularly, such as the "popular now" column on Bing.com, "trending topics" on Twitter.com, "trending now" on Yahoo.com, Google Trends, and so forth. Often popular topics are in the form of a list of keywords or phrases[1]. Take Twitter.com as an example. Clicking on a trending topic phrase will return a set of relevant Twitter posts (tweets) or web pages. Nonetheless, whether this is a convenient way for users to navigate through the popular topic information is still arguable. For example, when "SXSW" was listed as a trending topic, it seems difficult to understand at the first glance. A condensed topic summary would be extremely helpful for the users before diving into the massive search results to figure out what this topic phrase is about and why it is trending. In this paper, our goal is to generate a short text summary for any given topic phrase. Note that the proposed approach is not limited to trending topics, but can be applied to arbitrary Twitter topics.

There are a lot of differences between tweets and traditional written text that has been widely used for automatic summarization. In Table 1, we show example tweets for the topic "SXSW". The tweets were extracted by searching the Twitter site using the topic phrase as a query. We also provide an excerpt of the linked web content to help understand the topic. The tweets present some unique characteristics:

- All tweets are limited to 140 characters. Some tweets are news headlines from the official media, others are generated by users with various degrees of familiarity with the social media. The resulting tweets can be very different regarding the text quality and word usage.

[1]They are referred to as topic phrases hereafter, with no distinction between keywords and key phrases.

Proceedings of the Workshop on Language in Social Media (LSM 2011), pages 66–75,
Portland, Oregon, 23 June 2011.

Twitter Topic: "SXSW"	
Twts	I wish I could go to SXSW... I will, one day! *http://sxsw.com/*
	RT @user123: SXSW Film Round-Up: Documentaries *http://bit.ly/fg033b*
	@user456 yo.whats good,i met u at sxsw, talkin bout that feature.I was gonna see about sending u a few beats.u lookin for only original?
Web Cont	The South by Southwest (SXSW) Conferences & Festivals offer the unique convergence of original music, independent films, and emerging technologies...*(http://sxsw.com/)*

Table 1: Example tweets and an excerpt of the linked web content for Twitter topic "SXSW".

- Tweets lack structure information, contain various ill-formed sentences and grammatical errors. There are lots of noisy nonstandard tokens, such as abbreviations ("feelin" for "feeling"), substitutions ("Pr1mr0se" for "Primrose"), emoticons, etc.

- Twitter invented its own markup language. "@user" is used to reply to a specific user or call for attentions. The hashtag "#topic" aims to assign a topic label to the tweet, and is frequently employed by the twitter users.

- Tweets frequently contain embedded URLs that direct users to other online content, such as news web pages, blogs, organization homepages (Wu et al., 2011). According to Twitter's news release in September 2010 (Rao, 2010), 25% of tweets contain an URL. These linked web pages provide a much richer source of information than is possible in the 140-character tweet.

These Twitter-specific characteristics may pose challenges to the automatic summarization systems for identifying the essential information. In this paper, we focus on two such characteristics that are not studied in previous literature, the web content link and the non-standard tokens in tweets. Specifically, we ask two questions: (1) Is the web content linked from the tweets useful for summarization? Can we integrate different text sources, including the tweets and linked web pages, to generate more informative Twitter topic summaries? (2) what is the effect of nonstandard tokens on summarization performance? Will the summaries be improved if the noisy tweets were pre-normalized into standard English sentences? We investigate these two questions under a concept-based summarization framework using integer linear programming (ILP). We utilize text input that has various quality and is originated from multiple sources, and thoroughly analyze the resulting summaries using both automatic and human evaluation metrics.

2 Related Work

There is not much previous work on summarizing the Twitter topics. Most previous summarization literature focused on the written text domain, as driven by the annual evaluation tracks of the DUC (Document Understanding Conference) and TAC (Text Analysis Conference). To some extent, Twitter topic summarization is related to spoken document summarization, since both tasks deal with the conversational text that is contributed by multiple participants and contains lots of ill-formed sentences, colloquial expressions, nonstandard word tokens or high word error rate, etc. To summarize the spoken text, (Zechner, 2002) aimed to address problems related to disfluencies, extraction units, cross-speaker coherence, etc. (Maskey and Hirschberg, 2005; Murray et al., 2006; Galley, 2006; Xie et al., 2008; Liu and Liu, 2010a) incorporated lexical, structural, speaker, and discourse cues to generate textual summaries for broadcast news and meeting conversations.

For microblog summarization, (Sharifi et al., 2010a) proposed a phrase reinforcement (PR) algorithm to summarize the Twitter topic in one sentence. The algorithm builds a word graph using the topic phrase as the root node; each word node is weighted in proportion to its distance to the root and the corresponding phrase frequency. The summary sentence is selected as one of the highest weighted paths in the graph. (Sharifi et al., 2010b; Inouye, 2010) introduced a hybrid TF-IDF approach to extract one- or multiple-sentence summary for each topic. Sentences were ranked according to the average TF-IDF score of the consisting words; top weighted sentences were iteratively extracted, but excluding those that have high cosine similarity with the existing summary sentences. They showed the Hybrid TF-IDF approach performs constantly bet-

ter than the PR algorithm and other traditional summarization systems. Our approach of summarizing the Twitter topics is different from the above studies in that, we focus on exploring richer information sources (such as the online web content) and investigating effect of non-standard tokens. There are also studies working on visualizing Twitter topics by identifying a set of topic phrases and presenting the related tweets to users (O'Connor et al., 2010; Marcus et al., 2011). Our proposed approach can be beneficial to these systems by providing informative topic summaries generated from rich text sources.

3 Data Collection

We collected 5,537 topic phrases and the reference topic descriptions by crawling the Twitter.com and WhatTheTrend.com simultaneously during the period of Aug 22th, 2010 to Oct 30th, 2010 (about 70 days). The Twitter API was queried every 5 minutes for the current top ten trending topics. For each of these topics, a search query was submitted to the Twitter Search API to retrieve only English tweets related to this topic. If any tweet contains embedded URLs linked to the other web pages, the contents of these web pages were retrieved. For each topic, we limit the maximum number of retrieved tweets to 5,000 and webpages to 100. An example is shown in Table 1 for a topic phrase, some related tweets, and an excerpt of the linked webpage. WhatTheTrend API provides short topic descriptions contributed and constantly updated by the Twitter users. There is also a manually assigned category tag for each topic phrase. We found the top categories among the collected topics are "Entertainment (29.26%)", "Sports (25.58%)", and "Meme (15.69%, pointless babble)". We divided the collected topics into two groups: the general topics (e.g., "Chilean miners", "MTV VMA") and the hashtag topics that start with the "#" (e.g., "#top10rappers", "#octoberwish").

To generate reference summaries for the Twitter topics, two human annotators were asked to pick the topic descriptions/sentences (collected from WhatTheTrend.com) that are appropriate and valuable to be included in the summary. This is performed on a selected set of 1,511 topics with both trending duration and number of tweets greater than our predefined thresholds. For each of the topic sentences, we ask the annotators to label its category: (1) the sentence is a general description of the topic; (2) the sentence is trying to explain why the topic is trending; (3) it is hard to tell the difference. Overall, the two annotators have good agreement (Kappa = 0.67) regarding whether or not to include a sentence in the summary. Among the selected summary sentences, 22.58% of them were assigned with conflicting purpose tags such as (1) or (2). To form a reference summary, we concatenate all the topic sentences selected by both annotators. Since some reference descriptions are simply repetition of others with very minor changes, we reduce the duplicates by iteratively removing the oldest sentences if all the consisting words are covered by the remaining sentence collection, until no sentence can be removed. On average, the reference summary for general and hashtag topics contains 44 and 40 words respectively.

4 Summarization System

For each of the topic phrases, our goal is to generate a short textual summary that can best convey the main ideas of the topic contents. We explore and compare multiple text sources as summarization input, including the user-contributed tweets, web contents linked from the tweets, as well as combination of the two sources. The concept-based optimization approach (Gillick et al., 2009; Xie et al., 2009; Murray et al., 2010) was employed for selecting informative summary sentences and minimizing the redundancy. Note that our focus of this paper is not developing new summarization systems, but rather utilizing and integrating different text sources for generating more informative Twitter topic summaries.

4.1 Concept-based Optimization Framework

Concept-based summarization approach first extracts a set of important concepts for each topic, then selects a collection of sentences that can cover as many important concepts as possible, while within the specified length limit. This idea is realized using the integer linear programming-based (ILP) optimization framework, with objective function set to maximize the sum of the weighted concepts:

$$\max \quad \sum_i w_i c_i$$

where c_i is a binary variable indicating whether the concept i is covered by the summary; w_i is the weight assigned to c_i.

We enforce two sets of length constraints to the summary: sentence- or word-based. Sentence constraint requires the total number of selected summary sentences to not to exceed a length limit L_1; while word constraint requires the total words of selected sentences not to exceed length limit L_2. These two constraints are:

$$\sum_j s_j < L_1 \quad or \quad \sum_j l_j s_j < L_2$$

where s_j is a binary variable indicating whether sentence j was selected in the summary; l_j represents the number of words in s_j.

Further, we connect concept i with sentence j using two sets of constraints. For all the sentences that contain concept i, if any sentence was selected in the summary, the concept i should be covered by the summary; reversely, if concept i was covered by the summary, at least one of the sentences containing concept i should be selected.

$$\forall i \quad c_i \leq \sum_j o_{ij} s_j$$

$$\forall i, j \quad c_i \geq o_{ij} s_j$$

where the binary variable o_{ij} is used to indicate whether concept i exists in sentence j.

The concepts are selected by extracting n-grams (n=1, 2, 3) from the input documents corresponding to each topic. Similar to (Xie et al., 2009), we remove (1) n-grams that appear only once in the documents; (2) n-grams that have a consisting word with inverse document frequency (IDF) value lower than a threshold; (3) n-grams that are enclosed by higher order n-grams with the same frequency. These filters are designed to exclude insignificant n-grams from the concept set. The IDF scores were calculated from a large background corpus corresponding to the input text source, using individual sentences or tweets as pseudo-documents; words with low IDF scores (such as stopwords) tend to appear in many sentences and therefore should be removed from the concept set. We assign a weight w_i to an n-gram concept as follows:

$$w_i = tf(ngram_i) \times n \times \max_j \ idf(w_{ij})$$

where $tf(ngram_i)$ is the term frequency of $ngram_i$ in the input document of the topic; n denotes the order of $ngram_i$; w_{ij} are the consisting words of $ngram_i$; $idf(w_{ij})$ represents IDF value of word w_{ij}. This approach aims to extract n-grams that appear frequently in each topic, but do not appear frequently in a large background corpus. The weights are also biased towards longer n-grams since they carry more information.

4.2 Summarization Input

In this section, we explore different text sources as input to the summarization system. Different from previous studies that take input from a single text source, we propose to utilize both the user-contributed tweets and the linked web contents for Twitter topic summarization, since these two sources provide very different text quality and may contain complementary information regarding the topic. These text sources also pose great challenges to the summarization system: the tweets are short and extremely noisy; while the online contents linked from the tweets may have vastly different layouts and contain a variety of information.

4.2.1 Original Tweets

As shown in Table 1, the initially collected tweets are very noisy. They are passed through a set of preprocessors to remove non-ascii characters, HTML special characters, URLs, emoticons, punctuation marks, retweet tags (RT @user), etc. We also remove the reply (@) and hashtag (#) tokens that do not carry important syntactic roles (such as in the subject or object position) by using a set of regular expressions. These preprocessed tweets are sorted by date and taken as the first input source to the summarization system (denoted by "OrigTweets").

4.2.2 Normalized Tweets

The original tweets contain various nonstandard word tokens. In Table 2, we list the possible token categories and corresponding examples. We hypothesize that normalizing these nonstandard tokens into standard English words and using the normalized tweets as input can help boost the summarization performance.

We developed a twitter message normalization system based on the noisy-channel framework and a proposed letter transformation model (Liu et al.,

Category	Example
(1) abbreviation	tgthr, weeknd, shudnt
(2) phonetic sub w/- or w/o digit	4got, sumbody, kulture
(3) graphemic sub w/- or w/o digit	t0gether, h3r3, 5top, doinq
(4) typographic error	thimg, macam
(5) stylistic variation	betta, hubbie, cutie
(6) letter repetition	pleeeaas, togtherrr
(7) any combination of (1) to (6)	luvvvin, 2moro, m0rnin

Table 2: Nonstandard token categories and examples.

2011). Given a noisy tweet T, our goal is to normalize it into a standard English word sequence S. Under the noisy channel model, this is equivalent to finding the sequence $\hat{S}$ that maximizes $p(S|T)$:

$$\hat{S} = \arg\max_S p(S|T) = \arg\max_S (\prod_i p(T_i|S_i)) p(S)$$

where we assume that each non-standard token T_i is dependent on only one English word S_i, that is, we are not considering acronyms (e.g., "bbl" for "be back later") in this study. $p(S)$ can be calculated using a language model (LM). We formulate the process of generating a nonstandard token T_i from dictionary word S_i using a letter transformation model, and use the model confidence as the probability $p(T_i|S_i)$. This transformation process will be learned automatically through a sequence labeling framework. To form a nonstandard token, each letter in the dictionary word can be labeled with: (a) one of the 0-9 digits; (b) one of the 26 characters including itself; (c) the null character "-"; (d) a letter combination. We integrate character-, phonetic-, and syllable-level features in the model that can effectively characterize the formation process of nonstandard tokens. In general, the letter transformation approach will handle the nonstandard tokens listed in Table 2 yet without explicitly categorizing them. The proposed system also achieved robust performance using the automatically collected training word pairs. On a test set of 3,802 distinct nonstandard tokens collected from Twitter, our system achieved 68.88% 1-best normalization word accuracy and 78.27% 3-best accuracy.

We identify the nonstandard tokens that need to be normalized using the following criteria: (1) it is not in the CMU dictionary[2]; (2) it does not contain capitalized letter; (3) it appears infrequently in the topic (less than a threshold); (4) it is not a popular chat acronyms (such as "lol", "omg"); (5) it contains letters/digits/apostrophe, but should not be numbers only. These criteria are designed to avoid normalizing the named entities, frequently appearing out-of-vocabulary terms (such as "itunes"), chat acronyms, usernames, and hashtags. The selected nonstandard tokens in the original tweets will be replaced by the system generated 1-best candidate word. Note that we do not discriminate the context when replacing each nonstandard token. This will be addressed in the future work. We use these normalized tweets as a second source of summarization input and name them "NormTweets".

4.2.3 Linked Web Contents

For each Twitter topic, we collect a set of web pages linked by the topic tweets and use them as another source of summarization input. For each topic, we select up to n (n = 10) URLs that appear most frequently in the topic tweets and infrequently across different Twitter topics. This scheme is similar to the TF-IDF measure. This way we can select the salient URLs for each topic while avoiding the spam URLs. The contents of these URLs were collected and only distinct web pages were retained. We use an HTML parser[3] to extract the textual contents, and perform sentence segmentation (Reynar and Ratnaparkhi, 1997) on the parsed web pages. All the pages corresponding to the same topic were sorted by the date they were first cited in the tweets. These web pages were taken as another input text source for the summarization system, denoted as "Web".

4.2.4 Combining Tweets and Web Contents

We expect that taking advantage of both tweets and linked web contents would benefit the topic summarization system. Consolidating the distinct text sources may help boost the weight of key concepts and eliminate the spam information. As a preliminary study, we investigate concatenating either the original tweets or the normalized tweets with the linked web pages as input to the concept-based summarization system. This results in two inputs "Web + OrigTweets" and "Web + NormTweets". We will explore other ways of combining the two text

[2]http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[3]http://jericho.htmlparser.net/docs/index.html

sources in future work.

5 Experiments

5.1 Experimental Setup

Among the collected topics, we select 500 general topics (such as "Chilean miners") and 50 hashtag topics (such as "#octoberwish", "#wheniwasakid") for experimentation. On average, a general topic contains 1673 tweets and 3.43 extracted linked web pages; while a hashtag topic contains 3316 tweets but does not have meaningful linked web pages.

The concept-based optimization system was configured to extract a collection of sentences/tweets for each topic, using either the sentence- or word-constraint (denoted as "#Sent" and "#Word"). We opt to set individual length constraint for each topic rather than using a uniform length limit for all the topics, since the topics can be very different in length and duration. We use the number of sentences/words in the reference summary as the sentence/word constraint for each topic. Note that in practice this reference summary length information may not be available. We use the length constraints obtained from the reference summary in this exploratory study, since our focus is to first evaluate if twitter trending summarization is feasible, and what are the effects of different information sources and non-standard tokens. For a comparison to our approach, we implement the Hybrid TF-IDF approach in (Sharifi et al., 2010b; Inouye, 2010) as a baseline using "OrigTweets" as input. For the baseline, the summary length is altered according to the sentence- or word-constraint. The last summary tweet is cut in the middle if it exceeds the word limit.

The ROUGE-1 F-scores (Lin, 2004) are used to measure the n-gram (n=1) overlap between the system summaries and reference summaries. Since the ROUGE scores may not correlate well with the human judgments (Liu and Liu, 2010b), we also performed human evaluation by asking annotators to score both the system and reference summaries regarding the linguistic quality and content responsiveness, in the hope this will benefit future research in this direction.

5.2 Automatic Evaluation

We present the results (ROUGE-1 F-measure) for the general topics in Table 3. ROUGE-2 and ROUGE-4 scores show similar trends and thus are not presented. Five different text sources were exploited as the system inputs, as described in Section 4.2. To measure the quality of the input for summarization, we also include reference summary coverage score in the table, defined as the percentage of words in the reference summary that are covered by the input text source. When using tweets as input, we also investigate whether we should apply tweet normalization before or after the summarization process, that is "pre-normalization" (using "NormTweets" as input), or "post-normalization" (using "OrigTweets" as input, and rendering the normalized summary tweets).

General Topics		R-1 F(%)		RefSum
Input Source	Render	#Sent	#Word	Cov(%)
OrigTweets	Orig	29.53	30.21	94.81
	Norm	29.41	30.21	94.81
NormTweets	Norm	29.69	30.35	94.60
Web		24.32	25.07	63.74
Web + OrigTweets		29.58	30.44	95.37
Web + NormTweets		29.66	**30.54**	95.16
OrigTweets (Sharifi et al., 2010b)		24.37	25.68	94.81

Table 3: ROUGE-1 F-measure and reference summary coverage scores for general topics.

Compared to the Hybrid TF-IDF approach (Sharifi et al., 2010b; Inouye, 2010), our system performs significantly better ($p < 0.05$) according to the paired t-test; however, we also notice the ROUGE scores are lower compared to summarization in other text domains. This indicates that Twitter topic summarization is very challenging. Comparing the two constraints used in the concept-based optimization framework, we found that the word constraint performs constantly better for the general topics. This is natural since the word constraint tightly bounds the length of the system output, while the sentence constraint is relatively loose. For the different sources, we notice using linked web pages alone yields worse summarization performance, as well as lower reference summary coverage; however, when combined with the tweets, there is a slight increase in the coverage scores, and sometimes improved summarization results. This suggests that the linked web pages can contain extra

useful information for generating summaries. Regarding normalization, results show that the "pre-normalization" (using normalized tweets as input) can generally improve the summary tweet selection. For general topics, the best performance was achieved by combining the normalized tweets and linked web pages as input source and using the word-level constraint.

Hashtag Topics		R-1 F(%)		RefSum
Input Source	Render	#Sent	#Word	Cov(%)
OrigTweets	Orig	9.08	7.19	93.93
	Norm	9.09	7.16	93.93
NormTweets	Norm	**9.35**	7.14	93.71
OrigTweets (Sharifi et al., 2010b)		7.03	7.72	93.93

Table 4: ROUGE-1 F-measure and reference summary coverage scores for hashtag topics.

Results for hashtag topics were shown in Table 4 using tweets as input (there are no linked webpages for these topics). We notice the reference coverage scores are satisfying, yet the system output barely matches the reference summaries (very low ROUGE-1 scores). Looking at the reference and system generated summaries for the hashtag topics, we found the system output is more specific (e.g., "#octoberwish everything goes well."), while the reference summaries are often very general (e.g., "people tweeting about their wishes for October."). The human annotators also noted that most hashtag topics (such as "#octoberwish", "#wheniwasakid") are self-explainable and may require special attention to redefine an appropriate summary. Using sentence constraints yields better performance than word-based one, with larger performance difference than that for the general topics. We found the word-constraint summaries tend to include tweets that are very short and noisy. Our system with sentence-based length constraint also significantly outperforms the Hybrid TF-IDF approach (Sharifi et al., 2010b; Inouye, 2010). For hashtag topics, the best performance was achieved using the "pre-normalization" with sentence constraint.

For an analysis, we generate oracle system performance by using the reference summaries to extract a set of unweighted concepts to use in the ILP optimization framework for sentences/tweets selection. This results in 61.76% ROUGE-1 F-score for the general topics and 40.34% for the hashtag topics, indicating abundant space for future improvement. We also notice that though there is some performance gain using normalized tweets and linked web contents, the improvement is not statistically significant as compared to using the original tweets. Upon closer examination, we found the normalization system replaced 1.08% and 1.8% of the total word tokens for the general and hashtag topics respectively; these tokens spread in 13.12% and 16.85% of the total tweets. The relatively small percentage of the normalized tokens partly explains the marginal performance gain when using the normalized tweets as input. Similarly for linked web content, though it contains some sentences that can provide more details of the topic, but they can also take more space in the summary as compared to the short and condensed tweets. Therefore using the combined tweets and linked webpages does not significantly outperform using just the tweets.

5.3 Human Evaluation

	General			Hashtag	
	Tweet	Web	Ref	Tweet	Ref
Gram.	3.13	**3.42**	4.52	3.04	4.24
NRedun.	3.93	**4.64**	4.30	4.82	3.62
Clarity	**4.07**	3.91	4.77	4.06	4.60
Focus	**3.64**	3.03	4.75	3.22	4.72
Content	2.82	2.55	n/a	2.60	n/a
ExtraInfo	n/a	2.63	n/a	n/a	n/a

Table 5: Linguistic quality, content coverage, and usefulness scores judged by human assessors.

We ask two human annotators to manually evaluate the system and reference summaries regarding the readability and content coverage. Readability includes grammaticality, non-redundancy, referential clarity, and focus; content coverage was evaluated for system summaries against the reference summary. The annotators were also asked to rate the "Web" summaries regarding whether they provided extra useful topic information on top of the "Tweet" summary. 50 general topics and 25 hashtag topics were randomly selected for assessment. The "Tweet" and "Web" summaries were generated using the original tweets and linked web pages with word constraint for general topics, and sentence constraint for hashtag topics. Each of the assessors was

General Topic: "3PAR"	
RefSum	Dell Inc. and Hewlett-Packard Co. are both bidding for storage device maker 3Par Inc. 3Par jumped 21 percent after Hewlett- Packard Co. offered $30 a share for the company.
TweetSum	Dell ups 3Par offer yet again, to $27 per share Dell Raises 3par Offer to Match HP Bid Dell Matches HP's Offer for 3Par, Boosting Bid to $1.8 Billion
WebSum	Dell Matches HP's $27 Offer, Is Accepted by 3PAR. 3PAR has accepted an increased acquisition offer from Dell of US$27 per share, matching Hewlett-Packard's earlier raised bid.
Hashtag Topic: "#wheniwasakid"	
RefSum	when i was a kid.... people are sharing there best (good or bad) memories from childhood. People reminise the wonderful times about being a kid.
TweetSum	#whenIwasakid getting wasted meant eating all the ice cream and candy you could until you puked! #whenIWasAKid Apple & Blackberry were fruits not phones.

Table 6: Example system and reference summaries for both general and hashtag topics.

asked to judge all the summaries and assign a score for each criterion on a 1 to 5 Likert scale (5 being the best quality). The average scores of the two assessors were presented in Table 5.

For general topics, the "Web" summaries outperform the "Tweet" summaries on both grammaticality and non-redundancy, confirming the advantage of using the high-quality linked web pages. The referential clarity and focus scores of the "Web" summaries are not very high, since the summary sentences were extracted simultaneously from several web pages, and the system subjects to similar challenges as in multi-document summarization. The content coverage scores of both system summaries seem to correlate well with the ROUGE-1 F-measure, with a higher score for "Tweet" summaries. The assessors also rated that 48% of the "Web" summaries contain "Somewhat Useful" extra topic information, and 21% are "Very Useful". Note that this could be just because of the inherent difference of the two summaries, regardless of the input source, but in general we believe the linked web pages (such as the news documents) can provide more detailed and coherent stories as compared to the 140-character tweets. For hashtag topics, the "Tweet" summaries yield worse grammaticality and focus scores, but have very high non-redundancy score. On the contrary, the reference summaries often contain redundant information. The content match score between the system and reference summaries (2.6) does not seem to reflect the ROUGE scores. We hypothesize that even though the specificity of the two summaries is different, the assessors may still think the system summaries match the reference ones to some extent. A larger scale human evaluation is needed to study the correlation between human and automatic evaluation.

5.4 Discussions

We show an example of reference and system generated summaries for a general and a hashtag topic in Table 6, and summarize some challenges for this summarization task below:

- **Gold standard summaries are difficult and time-consuming to obtain.** The reference descriptions from WhatTheTrend.com were created by Twitter users, which vary a lot in word usage and would be unavoidably biased to the information available in Twitter. The user-contributed descriptions may also contain spam descriptions, repetitions, nonstandard tokens, etc. It would be better to have a concise non-redundant sentence collection for developing future summarization systems. In particular, hashtag topics need special attention. They account for 40% of the total trending topics in 2010 according to the statistics in WhatTheTrend.com[4]. Yet there still lacks standard definition regarding a good hashtag summary. From the example topic "#wheniwasakid" in Table 6, we can see they are very different in nature from general topics, thus future efforts are needed to define an appropriate summary.

[4]http://yearinreview.whatthetrend.com/

- **Evaluation issues.** Word based evaluation measures will rarely consider semantic relatedness between concepts, or name entity variations, such as "Hewlett-Packard" vs. "HP", "Dell ups 3Par offer" vs. "Dell Raises 3par Offer", etc. When comparing the system summaries with short human-written reference summaries, the word overlap varies a lot for different human summarizers.
- **Dynamically changing topics/events.** Some general topics are related to events that are constantly changing. Take the "3PAR" topic in Table 6 as an example, where two companies take turns to raise the bid for 3Par Inc. A good topic summary should be able to develop a series of sub-events and show the topic evolving process.

6 Conclusion

In this paper, we proposed to explore a variety of text sources for summarizing the Twitter topics. We employed the concept-based optimization framework with multiple input text sources to generate the summaries. We conducted both automatic and human evaluation regarding the summary quality. Better performance is observed when using the normalized tweets as input, indicating special treatment should be performed before feeding the noisy tweets to the summarization system. We also found the linked web contents can provide extra useful topic information. In future work, we will compare our system with other dedicated microblog summarization systems, as well as address some of the challenges identified in this study.

Acknowledgments

This work is partly supported by NSF award IIS-0845484. Any opinions expressed in this work are those of the authors and do not necessarily reflect the views of NSF.

References

Michel Galley. 2006. A skip-chain conditional random field for ranking meeting utterances by importance. In *Proc. of EMNLP*.

Dan Gillick, Korbinian Riedhammer, Benoit Favre, and Dilek Hakkani-Tür. 2009. A global optimization framework for meeting summarization. In *Proc. of ICASSP*.

David Inouye. 2010. Multiple post microblog summarization. *REU Research Final Report*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Workshop on Text Summarization Branches Out*.

Fei Liu and Yang Liu. 2010a. Exploring speaker characteristics for meeting summarization. In *Proc. of INTERSPEECH*.

Feifan Liu and Yang Liu. 2010b. Exploring correlation between ROUGE and human evaluation on meeting summaries. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(1):187–196.

Fei Liu, Fuliang Weng, Bingqing Wang, and Yang Liu. 2011. Insertion, deletion, or substitution? Normalizing text messages without pre-categorization nor supervision. In *Proc. of ACL-HLT*.

Adam Marcus, Michael S. Bernstein, Osama Badar, David R. Karger, Samuel Madden, and Robert C. Miller. 2011. TwitInfo: Aggregating and visualizing microblogs for event exploration. In *Proc. of CHI*.

Sameer Maskey and Julia Hirschberg. 2005. Comparing lexical, acoustic/prosodic, structural and discourse features for speech summarization. In *Proc. of Eurospeech*.

Gabriel Murray, Steve Renals, Jean Carletta, and Johanna Moore. 2006. Incorporating speaker and discourse features into speech summarization. In *Proc. of HLT-NAACL*.

Gabriel Murray, Giuseppe Carenini, and Raymond Ng. 2010. Interpretation and transformation for abstracting conversations. In *Proc. of NAACL*.

Brendan O'Connor, Michel Krieger, and David Ahn. 2010. Tweetmotif: Exploratory search and topic summarization for twitter. In *Proc. of the International AAAI Conference on Weblogs and Social Media*.

Leena Rao. 2010. Twitter seeing 90 million tweets per day, 25 percent contain links. *http://techcrunch.com/2010/09/14/twitter-seeing-90-million-tweets-per-day/*.

Jeffrey C. Reynar and Adwait Ratnaparkhi. 1997. A maximum entropy approach to identifying sentence boundaries. In *Proc. of the Fifth Conference on Applied Natural Language Processing*.

Beaux Sharifi, Mark-Anthony Hutton, and Jugal Kalita. 2010a. Summarizing microblogs automatically. In *Proc. of HLT/NAACL*.

Beaux Sharifi, Mark-Anthony Hutton, and Jugal K. Kalita. 2010b. Experiments in microblog summarization. In *Proc. of IEEE Second International Conference on Social Computing*.

Shaomei Wu, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. 2011. Who says what to whom on twitter. In *Proc. of WWW*.

Shasha Xie, Yang Liu, and Hui Lin. 2008. Evaluating the effectiveness of features and sampling in extractive meeting summarization. In *Proc. of IEEE Workshop on Spoken Language Technology*.

Shasha Xie, Benoit Favre, Dilek Hakkani-Tür, and Yang Liu. 2009. Leveraging sentence weights in a concept-based optimization framework for extractive meeting summarization. In *Proc. of INTERSPEECH*.

Klaus Zechner. 2002. Automatic summarization of open-domain multiparty dialogues in diverse genres. *Computational Linguistics*, 28(4):447–485.