Analysis of Tweet Generation as an Extractive Summarization Problem

First Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
`email@domain`

Second Author
Affiliation / Address line 1
Affiliation / Address line 2
Affiliation / Address line 3
`email@domain`

Abstract

Write abstract

1 Introduction

Write introduction
* Lot of work in summarization, twitter is a way of spreading data.
* about twitter
* tweets link to articles, classify as extractive summarization problem -¿ does it make sense? cite paper
* No, it doesn't.

Twitter is a public message broadcasting service with the constraint on the message being under 140 charachters. Since all posts on the website are public and a limited length, we can say that it qualifies as an altogether different genre of text - one that conveys precise message of a given topic in a limited number of charaters. This message can be an opinion about some happening or the news of it, and so on. The nature of this website has made it popular for global discussions on current goings-on in the world, with a large number of people constantly tweeting about all various topics.

2 Background and Related Work

Reduce this and add new papers
* summarization evaluation papers
* papers on classifying tweets based on intent?
* Add some glue at beginning of section

There has been work on using user ratings prediction for stylistic surface realisation Dethlefs et al. (2014). The study used ratings by users for generated texts along three axes of style, colloquialism, naturalness and politeness. The study then clustered users according to their ratings, and used stylistic predictions from this cluster towards the surface realization of new text. The three axes chosen were fairly arbitrarily, and may not have been completely independent. However, the concept of rating documents according to the stylistic characteristics of the text and using this stylistic information to rate the newly generated text is something that can be explored further.

Brooke et al. (2012) describe the process of building a formality lexicon by analyzing the stylistics of text. They calculate formality scores for words and sentences by training a model on a large corpus based on the appearance of words in specific documents. Their model represents words as vectors and the formal and informal seeds appear in opposite halves of the graphs, suggesting that we can use these seeds to determine if an article is formal or informal. Brooke and Hirst (2013) used an LDA based model using a similar idea of seed words for getting stylistic rankings for documents. The documents were ranked for styles such as literary, colloquial, subjective, concrete, and so on.

There have also been studies specific to Twitter data, for classifying and summarizing text, intents, etc. Ghosh et al. (2011) classified the retweeting activity of users based on entropy. The study considered the occurrence of the same URL in a different tweet as a retweet, and was able to separate the tweets as automatic or robotic retweeting, campaigns, news, blogs and so on. The study shows some interesting trends of retweeting activity for each of these cases. In another study,Chen et al. (2012), were able to extract sentiment expressions based on their corpus of tweets, that resulted in extraction of both formal and slang sentiment bearing words.

Mirco-blogging sites, easy access to Internet and the popularity of social media offers an opportunity to analyze data that comprises of statements from a huge number of users. Twitter is such a platform and has gained millions of users by now, and is hugely popular platform now for announcements, voicing opinions, promotions and so on. This data has been used for event summarization studies. O'Connor et al. (2010) uses topic summarizations for a given search for better browsing. Chakrabarti and Punera (2011) generate an event summary by learning the event using a Hidden Markov Model over the tweets describing it. Wang et al. (2014) generate a coherent event summary by treating summarization as an optimization problem for topic cohesion. Inouye and Kalita (2011) compare multiple summariza-

More details on each paper?

tion techniques to generate a summary of multi-post blogs on Twitter.

There has also been an attempt at generating tweets, texts of 140 characters using different text summarization techniques by Lloret and Palomar (2013). Summarization systems were used to summarize texts to sentences and then were compared against each other, evaluated using the ROUGE metric for evaluation. The ROUGE-1, ROUGE-2 and ROUGE-L metrics were used and the tweets were compared against an ideal summary. ROUGE is better when used with multiple reference texts and is not meant to be used at the sentence level. Thus the evaluation is done using the unigram, bigram and longest common subsequence matching techniques used in ROUGE-1, 2 and L. None of these techniques evaluate the fluency of the text, which is generally not expected from extractive summarization.

To the best of our knowledge, only Lofi and Krestel (2012) aim at generating tweets based on data from documents related to the topic. The system proposed uses keyword extraction techniques to generate tweets containing links to the article, hashtags based on the topic from documents and summarized content of the document. The study does not give details of implementation or evaluation of the system. Moreover, after the hashtags and the url, the Twitter constraint of 140 characters leaves room for few words in the generated tweet.

3 Data Extraction and Preprocessing

Reduce preprocessing part, add graphs

3.1 Using Twitter for Data Extraction

As mentioned earlier, there have been numerous studies that used data from the public Twitter feeds. However, since none of these datasets contained tweets and related articles promoted by these tweets separated into categories as required for this study, we extracted data directly from the site.

3.2 Extracting Data

Data was extracted from Twitter using the Twitter REST API using 51 search terms, or hashtags. These hashtags were chosen from a range of topics including pop culture, international summit meetings discussing political issues, lawsuits and trials, social issues and health care issues. All these hashtags were trending (being tweeted about at a high rate) at the time of extraction of the data. To give the data some variety, the data was extracted over the course of 15 days in November, which gave us multiple news stories to choose from for the search terms. A few examples of the search terms are shown in *Table 1* Only English tweets were extracted since the study is limited to English. In the beginning, about 30,000 tweets were extracted, and more than half of these tweets, around 16,000 contained URLs referencing some news articles, photos on photo sharing sites, and videos. The hashtags were chosen to maximise the number of articles related to the tweets. Hence, a lot of topics that were chosen were being tweeted about by news agencies and other popular news sources.

Can be compressed if required

Politics	Science & Technology
#apec2014 #G20 #oscarpistorius	#rosetta #lollipop #mangalayan
Events	**Films and Pop culture**
#haiyan #memorialday #ottawashootings	#TaylorSwift #theforceawakens #johnoliver
International	**Sports**
#berlinwall #ebola #erdogan	#ausvssa #playingitmyway #nycmarathon

Table 1: Table of Hashtags used for extraction. Table shows some examples of search terms chosen from various different categories.

The data from the tweets was cleaned by removing the tweets that were not in English as well as the ones that were retweeted, which is equivalent to re-publishing the same tweet from a different user.

Unique URLs were first extracted from the 16,000 or so URLs in the data. Next, data from these unique URLs was extracted and then preprocessed. The newspaper package was used to extract article text and the title from the web page. For the articles obtained from URLs, photos and video links for example, from Instagram and Youtube needed to be removed. For this, the data cleaning was achieved by removing articles by limiting word length of the extracted text to about 150 words. This ensured the removal of photos, videos, advertisements, incorrectly extracted articles from the data. After this preprocessing, the number of useful articles reduced from 6003 to 3066.

3.3 Current description of data

Add this part to Analysis?

The data currently consists of all tweets alongwith all the information of the tweet itself, such as the text of the tweet, links to articles if any, hashtags, and so on. The article links from these tweets are stored as a separate file, with information about the articles themselves, along with some preprocessed data. This includes the URL itself and the text extracted from the article, as well as some extracted information such as sentence boundaries, POS tags for tokens, parse trees and dependency trees. This processing of the text was done using the CoreNLP toolkit developed at Stanford (Manning et al., 2014 (Manning et al., 2014))

Tweets are linked to URLs through another file. A URL could have been tweeted through multiple tweets, all the ids of these tweets are linked to the same URL.

4 Analysis

Reduce

4.1 Subjectivity and Formality

After tagging a sample set of articles, the natural next task was to determine if the articles could be tagged automatically based on characteristics of the text. To achieve this, the degree of subjectivity and formality of the text was calculated with the help of some other studies. The subjectivity lexicon (Wilson et al., 2005 (Wilson et al., 2005)) was built using data for subjectivity analysis for a given text. The subjectivity lexicon consists of words that might indicate an opinion being expressed in a given text. Similarly, the formality lexicon gives was generated by Brooke et al. 2013 (Brooke and Hirst, 2013) and can be used to measure formality of a given text. The lexicon consists of words and phrases and the degree of formality for their occurrence. Thus, more formal words marked on a positive scale and informal words like those occurring in colloquial language are marked on a negative scale. Using the formality and subjectivity lexicons, the degree of subjectivity and formality of each individual article was calculated.

The degree of subjectivity returned a count per of the number of words present in the article that suggested an opinion per article. This number was normalized with the length of the article, and the degree of subjectivity was calculated per 10 words of an article. For this result, only the strong subjective entries in the lexicon were used to better differentiate between subjective and non-subjective articles.

The formality lexicon gave positive weights for formal expressions and negative for informal expressions. After calculating the formality weights for all articles, it was observed that they all had a total negative normalized weight, meaning a lot more informal expressions were getting matched. Hence, we used just the formal word occurrences for calculating the weight. Thus, above a certain cut-off weight, the article could be considered formal, else would be considered informal.

All the weights from both lexicons were averaged out over the articles relating to a single search term(or hashtag), and then ranked accordingly. The ranking showed that for subjectivity ranking over hashtags, films and music related hashtags are at the top, which would be the natural intuition given the nature of the topics. On the other hand, in the formality ranking, the hashtags relating to political issues had the highest formality ranking, while the hashtags for film titles, pop culture are all at the bottom. This also correlates with intuition about the topics. As a sanity check, we also looked at articles at the extreme points of the both the graphs. The texts of these articles suggested that they were consistent with the numbers.

Correlation between the rankings of hashtags given by both these experiments was calculated, and the Kendalls tau for this was 0.09 with a p-value of 0.34. The low correlation suggests that these two ways of evaluating subjectivity and formality are independent. The p-value suggests that there is not enough evidence to prove a correlation between subjectivity and formality of an article.

4.2 Correlating descriptive/non-descriptive with formal vs. informal for automatic tagging

To check if the descriptive vs non-descriptive tags correlated when tagged using the formality lexicon. If the document contained no formal words from the lexicon, it was tagged as non-traditional, else, it was tagged as traditional. The sample set of articles was tagged using this method, and after comparing them with the human tags, 42 out of the 62 tags matched, which gave a match percentage of 67%.

4.3 Position of tweet text in article experiments

4.3.1 Total match with text in article

We calculated the position of tweet text as a whole in the text. To compare the text, we removed the hashtags, references (@) and urls from the tweets. After this, we did direct substring comparison of the tweet in the text.

Out of the 6144 instances where a tweet text was checked against the text in the article, a complete match was found around 70 times. 30 times out of these, the tweet text had been matched against the title of the article extracted into the text. The rest of the results are significant, since the text of the tweet appears exactly as is inside the text of the article. The user who wrote the tweet for these articles went through the article text, and the sentence that either seemed to be the most conclusive contribution of the article, or expressed the opinion of the user were extracted to be tweeted.

We also checked to see if the tweet text matched a lot with the article titles, and this was found not to be the case. (*Needs to be verified)

4.3.2 Percentage match

Next, we did a percentage match with the text of the article after removing the stop words from both the tweet and the text. The results we got seem to suggest that a lot of significant words in the tweet are in fact present in the article. The minimum percentage match obtained was 60%.

4.3.3 Percentage matching inside a window in the article text

The next analysis was to check for a significant word matching inside a two or three sentence window inside the article text. We used a three sentence long window using the sentence boundary information obtained during preprocessing. After the text of the window was extracted, we performed a similar analysis as the last one, except on a smaller text. Next, the matching percentages from all such windows in the articles were compared and the maximum out of these was considered for the highest match percentage and match position for the final results. The final results are being verified, including the result for where the tweet text mostly comes from is random.

4.3.4 Least Common Subsequence match inside a window for the text

The percentage matched have mostly been a bag-of-words approach. The next step would be to look for phrases in the tweet coming directly from the text.

5 Evaluation

Perform and Write

6 Results

Write

7 Conclusion

Write

8 Future Work

write
* Classifying tweets based on intent, and being able to generate tweet might be generated from a template

Acknowledgments

References

Julian Brooke and Graeme Hirst. 2013. A multi-dimensional bayesian approach to lexical style. In *HLT-NAACL*, pages 673–679.

Julian Brooke, Vivian Tsang, David Jacob, Fraser Shein, and Graeme Hirst. 2012. Building readability lexicons with unannotated corpora. In *Proceedings of the First Workshop on Predicting and Improving Text Readability for target reader populations*, pages 33–39. Association for Computational Linguistics.

Deepayan Chakrabarti and Kunal Punera. 2011. Event summarization using tweets. *ICWSM*, 11:66–73.

Lu Chen, Wenbo Wang, Meenakshi Nagarajan, Shaojun Wang, and Amit P Sheth. 2012. Extracting diverse sentiment expressions with target-dependent polarity from twitter. In *ICWSM*.

Nina Dethlefs, Heriberto Cuayáhuitl, Helen Hastie, Verena Rieser, and Oliver Lemon. 2014. Cluster-based prediction of user ratings for stylistic surface realisation. *EACL 2014*, page 702.

Rumi Ghosh, Tawan Surachawala, and Kristina Lerman. 2011. Entropy-based classification of'retweeting'activity on twitter. *arXiv preprint arXiv:1106.0346*.

David Inouye and Jugal K Kalita. 2011. Comparing twitter summarization algorithms for multiple post summaries. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 298–306. IEEE.

Elena Lloret and Manuel Palomar. 2013. Towards automatic tweet generation: A comparative study from the text summarization perspective in the journalism genre. *Expert Systems with Applications*, 40(16):6624–6630.

Christoph Lofi and Ralf Krestel. 2012. iparticipate: Automatic tweet generation from local government data. In *Database Systems for Advanced Applications*, pages 295–298. Springer.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

Brendan O'Connor, Michel Krieger, and David Ahn. 2010. Tweetmotif: Exploratory search and topic summarization for twitter. In *ICWSM*.

Lu Wang, Claire Cardie, and Galen Marchetti. 2014. Socially-informed timeline generation for complex events. *constitution*.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics.