

EXAMINING INDICATIVE TWEET GENERATION AS AN
EXTRACTIVE SUMMARIZATION PROBLEM

by
Priya Sidhaye

School of Computer Science
McGill University, Montréal
2016

A THESIS SUBMITTED TO THE FACULTY OF GRADUATE STUDIES AND RESEARCH
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE

© Copyright 2016 by Priya Sidhaye

Abstract

Social media such as Twitter have become an important method of communication, with potential opportunities for Natural Language Generation (NLG) to facilitate the generation of social media content. The presence of the URL in the short tweet is a strong signal that the tweet is functioning to help Twitter users decide whether to read the full article. We focus on the generation of these *indicative tweets*, which we define as tweets containing links to external web pages. Extractive summarization is the process of generating a summary of the text by choosing a subset of words from the text. Previous work in automatic tweet generation has viewed the linked web page as the source text from which the tweet is generated in an extractive summarization setting. However, this setting may not be appropriate, because it is unclear to what extent indicative tweets actually behave like extractive summaries.

We collect a corpus of indicative tweets with their associated articles and investigate the extent to which they can be derived from the articles using extractive methods. We also consider the impact of the formality and genre of the article. We conduct further studies to detect the *function* of the tweet, i.e., the reason the user shared the tweet. In particular, with the aim of finding possible factors that influence the composition of the tweet, we considered whether the tweet is an advertisement or summary of the article. We find a significant positive correlation between the degree of extraction and the degree to which a tweet behaves like a summary.

Our results demonstrate the limits of viewing indicative tweet generation as extractive summarization, and point to the need for the development of methods for tweet generation that are not based on simple extractive techniques. We also show the need for genre-sensitive methods

for generating tweets.

This thesis contributes a novel dataset of indicative tweets labelled by their function as defined above. The tweets are categorized by their subjects, and span broad areas of public interest. In terms of theoretical contributions, this thesis clarifies the relationship between an indicative tweet and the article to which it relates and lays the groundwork for developing a system that would be able to generate indicative tweets from referenced documents, possibly based on parameters such as whether the tweet is intended to be an advertisement or a public service announcement.

Résumé

Les réseaux sociaux tels que Twitter sont devenus un moyen de communication important, soulignant de plus en plus de possibilités de génération automatique de texte (NLG) pour faciliter la génération des textes qu'on rencontre dans les réseaux sociaux. Nous nous concentrons sur la génération de *tweets indicatifs* qu'on définit comme étant des tweets contenant un lien pointant vers de pages web externes. Le résumé automatique par extraction consiste à générer un résumé du texte en sélectionnant certains mots de ce texte. Les approches antérieures à la génération automatique de tweets ont considéré que la page web liée est la source à partir de laquelle le tweet est généré dans un cadre de résumé automatique par extraction. Toutefois, ce cadre-là pourrait ne pas être le plus convenable puisqu'il n'est pas assez évident dans quelle mesure les tweets indicatifs se comportent comme étant des résumés par extraction.

Nous collectons un corpus de tweets indicatifs avec leurs articles associés et étudions dans quelle mesure ils peuvent être obtenus à partir des articles en utilisant des méthodes de résumé automatique par extraction. Nous considérons également l'effet de la formalité et le genre de l'article. Nous menons des études pour spécifier et mieux comprendre la *fonction* du tweet, c'est-à-dire, la raison pour laquelle l'utilisateur a partagé le tweet. Particulièrement, dans le but de trouver les facteurs qui agissent sur la composition du tweet, nous considérons si le tweet est une annonce publique ou un résumé de l'article. Nous trouvons une corrélation positive considérable entre le degré d'extraction et le degré selon lequel le tweet se comporte comme étant un résumé. Nos résultats démontrent les limites de considérer la génération des tweets

indicatifs comme étant une méthode de résumé automatique par extraction. Ils soulignent également le besoin de développer de méthodes de génération de tweets qui ne sont pas basées sur de simples techniques d'extraction. Nous démontrons de même le besoin de méthodes de génération de tweets qui sont sensibles au genre.

Cette thèse présente un nouvel ensemble de données de tweets indicatifs que nous avons classé par leurs fonctions comme définies ci-dessus. Les tweets sont classés par leurs sujets et couvrent de grands domaines d'intérêt public. En tant que contribution théorique, cette thèse clarifie la relation entre un tweet indicatif et l'article correspondant et présente un plan pour le développement d'un système qui pourrait générer des tweets indicatifs à partir de documents référencés, en se basant sur de paramètres qui peuvent, par exemple, considérer si le tweet fonctionne comme une publicité ou un message d'intérêt public.

Acknowledgements

I would like to thank my supervisor, Jackie Cheung, who has been a teacher and advisor and helped me not just by guiding my thesis, but also by creating a fun research group that I am proud to be a part of.

I would like to thank Prof. Joelle Pineau for putting me in touch with Jackie for my thesis, Jad Kabbara for helping me translate my abstract to French, and everyone in the very large Reasoning and Learning lab family.

I would also like to thank my parents, my family and friends who have been a constant emotional support system, to be always relied upon.

I would like to thank Julian Brooke for providing the formality lexicon used in part of this study, Prof. Derek Ruths his comments on the thesis, and Prof. Joelle Pineau for agreeing to be the external reviewer of the thesis.

Last but not the least, I thank the McGill University community for a wonderful experience and for teaching me a thing or two beyond the subjects I came here to study.

Contents

1	Introduction	1
2	Background and Related Work	5
2.1	Summarization	5
2.1.1	ROUGE: Evaluation Measure for Text Summarization	7
2.2	Studies Based on Twitter Data	8
2.2.1	Twitter Data and Summarization	8
2.2.2	Classifying Twitter Data	9
3	Data Extraction and Preprocessing	11
3.1	The Need for a New Dataset	11
3.2	Extracting Data	12
4	Analysis of Dataset for Determining Relationship between Tweet and Article	15
4.1	Exact Match Calculations	17
4.2	Percentage Match for Unigrams	18
4.3	Percentage Match for Bigrams	20
4.4	Percentage Match Inside a Window in the Article Text	22
4.5	Longest Common Subsequence Match	24
4.6	Interaction with Formality	25
4.6.1	Examining Formality Scores with Respect to Match Percentages	26

5	User Study for Identifying Functions of Tweets	28
5.1	Functions of Tweets	29
5.2	User Study Design	29
5.2.1	Questions Used in the User Study	29
5.2.2	Running the User Study	31
5.2.3	Qualification Details	31
5.2.4	Pilot Studies	36
5.3	Results and Analysis for the User Study	37
5.3.1	Conclusions from the User Study	41
6	Conclusion	43
6.1	Future Work	44
6.1.1	Study Functions and Intents	44
6.1.2	A Structure for Generating Tweets	44
6.1.3	Parameterized summarization	44
	Bibliography	45

List of Tables

3.1	Hashtags used for extraction, grouped into various categories.	12
3.2	Example of a tweet, title of the article and the text.	13
4.1	An example of a tweet, the title of the article it links to, and the text of the article, where the tweet cannot be extracted from the text.	16
4.2	An example of a tweet, the title of the article it links to, and the text, where the tweet can be extracted from the text. The matched portions of the tweet and article are in bold.	16
4.3	Measures and sections they are found in	17
4.4	List of stop words in NLTK	17
4.5	Example where tweet is extracted as is from the text (matched portion in bold).	18
4.6	Order of formality ranking in hashtags	26
4.7	Example of formality in article affecting tweet	27
5.1	Questions used in user study	30
5.2	Analysis of user study results	38
5.3	Example where all three workers said it was not an advertisement.	38
5.4	Example where all three workers said the tweet was an advertisement for article.	38
5.5	Example where all three raters said the tweet was not a summary.	38
5.6	Example where all three raters agreed the tweet was a summary.	38

5.7	Mann Whitney U test results for the advertisement question(indicativeness):	
	Unigram Match	39
5.8	Mann Whitney U test results for the summary question(informativeness): Un-	
	igram Match	39
5.9	Mann Whitney U test results for the advertisement question(indicativeness):	
	Bigram Match	39
5.10	Mann Whitney U test results for the summary question(informativeness): Bi-	
	gram Match	40
5.11	Mann Whitney U test results for the advertisement question(indicativeness):	
	Longest Common Subsequence	40
5.12	Mann Whitney U test results for the summary question(informativeness): Longest	
	Common Subsequence	40

List of Figures

2.1	Examples of extractive and abstractive summarization. The extractive summary is highlighted by the blue box, while the abstractive summary is highlighted by the red box.	6
4.1	Unigram matching percentages	19
4.2	Histogram for absolute number of unigrams matched	20
4.3	Bigram match percentages	21
4.4	Histogram for absolute number of matched bigrams.	21
4.5	Match percentages in tweet against window in article	23
4.6	LCS match percentages	24
5.1	User study question 1 example	32
5.2	User study question 2 example	34

Chapter 1

Introduction

Social media comprise a large part of our lives, with various outlets providing a platform for sharing thoughts, news, images and videos. With the rise in popularity of social media, message broadcasting sites such as Twitter and other microblogging services have become an important means of communication, with an estimated 500 million tweets being written every day¹. In addition to individual users, various organizations and public figures such as newspapers, government officials and entertainers have established themselves on social media in order to disseminate information or promote their products. Social media thus provide an incredibly dense and varied source of data, originating from people and organizations all over the world.

Following the prolific increase in the use of social media, there has been an increase in the number of studies in natural language processing using the sources of data social media provide. Specifically for Twitter, these include areas such as tweet parsing (Ritter et al., 2011; Kong et al., 2014), text normalization (Han and Baldwin, 2011; Kaufmann and Kalita, 2010), tweet POS tagging (Gimpel et al., 2011; Owoputi et al., 2013), sentiment analysis of tweets (Kouloumpis et al., 2011; Mohammad et al., 2013b), event summarization (Chakrabarti and Punera, 2011; Nichols et al., 2012), identifying bot behaviours (Chu et al., 2012) and inferring things like political views of individuals (Mohammad et al., 2013a). While this progress in the

¹<https://about.twitter.com/company>

development of Twitter-specific POS taggers, parsers, and other tweet understanding tools is encouraging, there has been little work on methods for *generating tweets*. Methods to generate tweets would be beneficial to users and organizations for the purposes of advertisement, education, or even entertainment. Examples of such uses include generating tweets that advertise products or services based on some online review articles, or notifying users of closure of roads because of construction work by local governments.

In this thesis, we study the generation of the particular class of tweets that contain a link to an external web page that is composed primarily of text. Given the short length of a tweet, the presence of a URL in the tweet is a strong signal that the tweet is functioning to help Twitter users decide whether to read the full article. We call this class of tweets *indicative tweets*, since they act as indicative summaries of the articles they are being linked to. Indicative tweets represent a large subset of tweets overall, constituting more than half (53.4%) of the tweets in a data set that we collected in Chapter 3. Generating indicative tweets would appear to be a feasible problem to solve using current methods in text summarization, such as extractive summarization, because there is a clear source of input from which a tweet could be generated.

There has in fact been some work along these lines, within the framework of extractive summarization. Lofi and Krestel (2012) describe a system to generate tweets from local government records through keyword generation. However, they do not provide a formal evaluation for their proposed system.

Lloret and Palomar (2013) compare various extractive summarization algorithms applied on Twitter data to generate tweets from documents. They compare the overlap between system-generated and user-generated tweets using ROUGE (Lin, 2004a), a recall-based evaluation metric for summarization, and achieve some success in generating tweets based on ROUGE scores. Unfortunately, they also show that there is little correlation between ROUGE scores and the perceived quality of the tweets when rated by human users for indicativeness and interest. An *indicative* text is one that aims to point to or generate interest about something. Hence the *indicativeness* of a tweet can be defined as a measure of how strongly it points to the

article. More discussion about these studies is done in Chapter 2.

Beyond issues of evaluation measures, it is also unclear whether extraction is the strategy employed by human tweeters. One of the original motivations behind extractive summarization for news text was the observation that human summary writers tended to extract snippets of key phrases from the source text (Mani, 2001). And while it may be true that an automatic tweet generation system need not necessarily follow the same approach to writing as human tweeters, it is still necessary to know what proportion of tweets could be accounted for in an extractive summarization paradigm. More scrutiny is required to determine whether methods and evaluation schemes from extractive summarization can be adopted for the purpose of producing indicative tweets and is one of the primary aims of this thesis. With indicative tweets, an additional issue arises in that the genre of the source text is not constrained; for example it may be a news article or an informal blog post or an advertisement. This genre of the source text may be vastly different from the desired formality of tweet itself, and thus, a genre-appropriate extract may not be available.

Contributions We begin to address the above issues through a study that examines to what extent tweet generation can be viewed as an extractive summarization problem. We extracted a dataset of indicative tweets containing a link to an external article, including the documents linked to by the tweets. We used this data and applied unigram, bigram and LCS (longest common subsequence) matching techniques inspired by ROUGE to determine what proportion of tweets can be found in the linked article. This measure can also be defined as *extractiveness* of the tweet, or the degree to which the tweet has been extracted from the article. Even with the permissive unigram match measure, we find that well under half of the tweet can be found in the linked article. We also use stylistic analysis on the articles to examine the role that genre differences between the source text and the target tweet play and if genre can give an indication for whether the tweet can be extracted. We find that tweets are extracted from articles with higher formality to a greater extent than ones with lower formality.

We further conducted studies to identify functions for the tweets with respect to the articles. The data extracted from Twitter was presented to workers on a crowdsourcing website, to ask whether the tweets were indicative or informative. *Informative* tweets are the ones that convey some information from the article, and *informativeness* is defined as the degree to which the information from the article is conveyed in the tweet. We found a link between whether the tweet was deemed informative, and the degree to which a tweet has been extracted from the article, offering a better view of our ROUGE-inspired analysis methods detailed in Chapter 4 and when they can be used for generating tweets. As a result, this dataset tagged by human evaluators has been generated and should be useful in further studies for identifying functions of tweets.

Overall, our results point to the need for the development of a methodology for indicative tweet generation, rather than to expropriate the extractive summarization paradigm that was developed mostly on news text. Such a methodology will ideally be sensitive to stylistic factors as well as the underlying intent of the tweet.

Chapter Outline Chapter 2 contains the discussion on various related studies. The process of collecting the dataset is detailed in Chapter 3. Chapter 4 contains the analyses performed on the dataset we collected. Chapter 5 details the process of designing and executing the user study based on our data and analysis of the input from the workers. Finally, Chapter 6 contains the conclusions drawn from our the discussion of results in Chapter 4 and Chapter 5.

Portions of Chapter 1, Chapter 3 and Chapter 4 were published as part of the conference paper Sidhaye and Cheung (2015). The contribution of the co-author was that of a thesis supervisor.

Chapter 2

Background and Related Work

This chapter surveys various concepts used in the thesis, including an introduction to automatic summarization, methods for evaluating summaries, as well as studies relating to Twitter data and tweet generation.

2.1 Summarization

Text summarization is the task of condensing an original text document or documents while retaining as much of the important information as possible (Mani, 2001). In addition, the summary must also satisfy goals related to the quality of the generated text, such as readability and coherence. The two main approaches for automatic text summarization are *extractive* and *abstractive* summarization.

Extractive summarization uses the technique of choosing the important parts of the text and rearranging them to generate summaries. Nenkova and McKeown (2012) describe the components in extractive summarization techniques as 1) building an internal representation of the important parts of the text, 2) ranking these in the order of importance or time or other relevant metric based on context, and then 3) selecting a suitable list of these sentences to form the summary.

This selection of important parts of the source can be done at various levels, such as phrases,



Figure 2.1: Examples of extractive and abstractive summarization. The extractive summary is highlighted by the blue box, while the abstractive summary is highlighted by the red box.

sentences, or paragraphs (Nenkova and McKeown, 2012; Hahn and Mani, 2000). In phrase-level summarization, smoothing techniques may be used to generate readable texts, since stitching together phrases from the source will lack coherence. In contrast to phrase-level summarization, sentence-level summarization techniques tend to be inherently more grammatically correct since sentences are directly picked out. However, sentence compression techniques can also be used to reduce the size of the summaries in sentence-level summarization (Knight and Marcu, 2002). Even after using smoothing techniques to generate readable text, extractive summaries tend to be incoherent and hard to read (Liu and Liu, 2009). It should be noted that since we are dealing with the summaries being used as tweets, and since tweets are only 140 characters long, we will mostly be dealing with word-level or n-gram-level summarization.

The second approach is that of abstractive summarization. This is a summarization approach that aims to keep the content or meaning of the input source the same while condensing the text or generalizing it, and involves text generation for generating the summary. As a rule, abstractive summarization requires world knowledge and is a much more difficult problem to solve. As a result, current summarization techniques concentrate on improving results from extractive summarization (Nenkova and McKeown, 2012).

Figure 2.1 shows examples of both an extractive and abstractive summary in the form of a newspaper article thumbnail. The blue box outlines a few sentences from the article that have been picked to give a brief description. This acts as the extractive summary. The red box out-

lines the title of the article, which is an abstractive summary of the article; it is a generalization of the events described, carefully omitting details yet leaving the overall meaning of the event untouched.

Although extractive summarization has been predominant, there have been studies on its limitations. He et al. (2000) compared user preferences for various mechanisms of browsing content from an audio-visual presentation. They demonstrated that the most preferred method of summarization was highlights and notes provided by the author, rather than transcripts or slides from the presentation, which can be viewed as the full source text and the compressed pointers for the presentation respectively. Conroy et al. (2006) investigated the issue of limits of extraction by using an oracle ROUGE score based on a probabilistic model of unigrams that might appear in the gold standard summaries and exploit this to create a new method of summarization that uses maximum likelihood estimation.

2.1.1 ROUGE: Evaluation Measure for Text Summarization

ROUGE(Recall-Oriented Understudy for Gisting Evaluation) is an evaluation measure popularly used for evaluating the quality of summaries (Lin, 2004b). It measures the quality of a summary by comparing the output of the system being tested against a set of gold standard summaries by word or n-gram overlap. The intuition is that if the generated summary has enough in common with a set of human-written summaries, then it can be judged as a good summary. The different types of comparisons calculated are the unigram, bigram, trigram and least common subsequence (ROUGE-1,2,3 and L respectively). A set of gold standard summaries are used to account for the fact that summary writers do not agree on the contents of the summary.

$$ROUGE-n = \frac{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)} \quad (2.1)$$

The equation as described by Lin and Och (2004) shows the calculation of the ROUGE-n

score, where n is the length of the n-gram $gram_n$, and $Count_{match}$ is the count of co-occurring n-grams in the candidate summary and the reference summary, S . The sum of $Count_{match}$ for all the reference summaries gives us the true positives for the recall based measure. The ratio of this sum and count of all n-grams in the reference summaries gives the final ROUGE-n score.

The use of multiple gold standard summaries gives rise to a subjective evaluation metric where the quality of the evaluation is dependent on the quality and number of the gold standard summaries. ROUGE also does not take into account whether the summary is fluent or coherent. However, ROUGE is useful for the evaluation of summarization methods for overall content retention from the original text. This is possible due to the use of n-gram co-occurrence statistics used by ROUGE.

2.2 Studies Based on Twitter Data

2.2.1 Twitter Data and Summarization

In this section, we discuss studies that use Twitter data as the source to study the summarization concepts discussed above.

The following studies focus on using summarization in relation to Twitter data. O'Connor et al. (2010) use topic summarization for a given search for better browsing. Chakrabarti and Punera (2011) generate an event summary by learning about 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 summarization techniques to generate a summary of multi-post blogs on Twitter. Wei and Gao (2014) use tweets to help in generating better summaries of news articles.

As described in Chapter 1, we analyze tweet generation using measures inspired by extractive summarization evaluation. Lloret and Palomar (2013) compared the different text

summarization techniques for tweet generation. Summarization systems were used to generate sentences which could then be taken to be tweets by summarizing documents to lengths smaller than 140 characters. The system-generated tweets were evaluated using ROUGE measures (Lin, 2004a). The ROUGE-1, ROUGE-2 and ROUGE-L measures were used, and a human-written reference tweet was taken to be the gold standard. However, they found that the generated tweets did not rank well when evaluated by humans, even though the tweets achieved success in terms of ROUGE scores.

The Lloret and Palomar (2013) study shows that extractive summarization algorithms may not generate good quality summaries despite giving high ROUGE evaluation scores. Cheung and Penn (2013) show that for the news genre, extractive summarization systems that are optimized for *centrality*—that is, getting the core parts of the text into the summary—cannot perform well when compared to model summaries, since the model summaries are abstracted from the document to a large extent. Since ROUGE is such an evaluation method, we can call into question the use of extractive summarization for tweet generation in the news genre. This question fits into the bigger problem of whether extractive summarization can be used for tweet generation, and is discussed further in the following chapters.

2.2.2 Classifying Twitter Data

In this section we discuss concepts which we consider in our user studies, such as why the tweet was written, and what purpose does it serve.

Ghosh et al. (2011) classified the retweeting activity of users based on time intervals between retweets of a single user and frequency of retweets from unique users. They defined ‘retweet’ as the occurrence of the same URL in a different tweet. The study was able to classify the retweeting as automatic or robotic retweeting, campaigns, news, and blogs, based on the time-interval and user-frequency distributions.

We define *function* of a tweet as why the user chose to write and share the tweet. This function can be considered on multiple levels. The closest description to our idea of function

was found in Sinclair and Ball (1996), who describe this as ‘communicative intent’, describing the types as information, discussion, recommendation, recreation, religion and instruction with further subcategories. This higher level function is described as *intent* in many studies, and is akin to classifying the topic or genre of the tweet. Examples of high-level function include indicative tweets, informative tweets or critical tweets. Lower-level functions of a tweet could be an advertisement for a car, an announcement of an event or drawing attention to a particular sentence in an article.

There are several studies on classifying the function of tweets. Wang et al. (2015) use bootstrapping to generate an intent keyword set used in generating an intent graph in a semi-supervised manner. They focus on finding tweets with intent and then classifying those tweets. The intent here is defined as a wish or a plan for some action, such as intent for buying/doing something such as food, drink, travel or career. Classification of intents in this way can directly be used as intents for purchasing and be utilized for advertisements. For example, if an intent for buying a new car is detected from a tweet by the user, advertisements of cars would be shown to the user. Banerjee et al. (2012) analyze real time data to detect presence of intents in tweets. Gómez-Adorno et al. (2014) use features from text and stylistics to determine user intentions, which are classified as news report, news opinion, publicity, general opinion, share location, chat, question or personal message. Mohammad et al. (2013a) take a different approach on the intents and use them to study the classification of user intents specifically for tweets related to elections. They study tweets related to one election and classify tweets as ones that agree or disagree with the candidate, or contain humour, support, sarcasm, or irony. They group these tweets into a broader classification of favouring vs opposing sentiments.

The studies discussed in this chapter lay the groundwork for the tasks performed in the next few chapters; namely, collecting the data, analyzing the data, and running user studies on the data. We now discuss the process of building the dataset in Chapter 3.

Chapter 3

Data Extraction and Preprocessing

This chapter discusses the need for a new dataset of indicative tweets and linked articles, the decision to use Twitter data for the study, and the process of data collection for the thesis. The process of data collection includes the methods used for extracting data from Twitter, and the preprocessing done on the data to prepare it for analysis.

3.1 The Need for a New Dataset

As mentioned earlier, there have been numerous studies that used data from the public Twitter feeds. However, only a few of the datasets in those studies focused on tweets in conjunction with articles linked to these tweets. One such dataset was collected by Lloret and Palomar (2013), but it only contains 200 English tweet-article pairs. Wei and Gao (2014) also constructed a dataset that contains both tweets and articles linked to by the tweets, but this data only deals with news text, and does not contain the variety of topics and genres we wanted in the data. These datasets did not give us the ability to explore relationships between amount of extraction in tweets, genres and other stylistic factors such as formality, as described in Chapter 2. We therefore chose to build our own dataset. The following section describes the extraction, cleaning and other preprocessing steps that we performed on the data.

Science & Technology	Entertainment	Events	Miscellaneous
#android #cometlanding #lollipop #lollipopupdate #mangalayan #nexus6 #philae #rosetta	#1989 #BBCSyriaWars #betterstarwarstitles #harrypotter #interstellar #johnoliver #moneyball #montythepenguin #TaylorSwift #theforceawakens #winteriscoming	#bahamas #buffalosnow #haiyan #lestweforget #MarysvilleShooting #memorialday #ottawashootings #snowstorm	#1wtc #abercrombieandfitch #annefrank #beenrapedneverreported #KevinVickers #mentalhealth #netneutrality #pointergate #RobertONeill
Politics	International	Sports	Legal
#apec #apec2014 #cdnpoli #G20 #GOP	#berlinwall #canadachinatradeddeal #ebola #erdogan #obamacare #putin #syria	#ausvssa #nycmarathon #playingitmyway	#ghomeshi #oscarpistorius

Table 3.1: Hashtags used for extraction, grouped into various categories.

3.2 Extracting Data

Data was extracted from Twitter using the Twitter REST API using 51 ‘hashtags’, which are handles with which tweets are tagged. 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 get a broader sample, the data was extracted over the course of 15 days in November, 2014, rather than on a single point in time. From this set of possible topics, we selected hashtags such that there would be broad representation in terms of various stylistic properties of text (e.g. formality and subjectivity), different genres, and a variety of sources of articles. Different kinds of sources would include established news outlets, blogs, and individuals. All the search terms used are shown in Table 3.1, and have been classified into different genres for the purpose of ease of reading.

We extracted 30,621 tweets, of which more than half, or 16,349, contained URLs to an

Tweet	#RiggsReport: #CA as the #ElectionNight exception. Voters rewarded #GOP nationally, but not in the #GoldenState. http://t.co/K542wvSNVz
Title	The Riggs Report: California as the Election Night exception
Text	When the dust settled on Election Night last week...

Table 3.2: Example of a tweet, title of the article and the text.

external news article, photo on a photo sharing site, or video. The hashtags were chosen to maximise the number of articles linked to by the tweets. Many topics that were chosen were being tweeted about by news agencies and other popular news sources.

Before extracting the URLs, The data from the tweets was cleaned by removing the tweets that were not in English. We also removed retweets; i.e., re-publications of a tweet by a different user.

We deduplicated the 16,349 extracted URLs into 6,003 unique addresses, then extracted and preprocessed their contents. The `newspaper` package¹ was used to extract article text and the title from the web page. Since we are interested in text articles that can serve as the source text for summarization algorithms, we needed to remove photos and video links such as those from Instagram and YouTube. To do so, we removed those links that contained fewer than a threshold of 150 words. This preprocessing reduced the number of useful articles from 6,003 to 3,066. Further tweet-article pairs where the text of the tweets was identical were removed and the number of remaining unique tweet-article pairs was 2471. It should be noted that this extraction tool is not very effective at extracting titles from text.

The final version of the data consists of tweets along with other information about the tweet: links to articles, hashtags, time of publication and other details provided by the Twitter API, which were not relevant to our analyses. We also retained the linked article text and preprocessed it using the CoreNLP toolkit developed by Manning et al. (2014). 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. These annotations are used later during our analysis in Chapter 4. Table 3.2 shows an example of an

¹<https://pypi.python.org/pypi/newspaper>

entry in the dataset. A URL could have been tweeted through multiple tweets: i.e., the `ids` of these tweets are linked to the same URL.

We will discuss the analyses performed on this dataset in Chapter 4.

Chapter 4

Analysis of Dataset for Determining Relationship between Tweet and Article

In the previous chapters, we have described the background and setup of the question at hand; that is, whether extractive summarization is an adequate solution for tweet generation. We have also described the creation of a dataset from Twitter that enables us to explore this exact question and glean more information about how stylistic factors might influence the answers. To answer this question, we now describe the analyses we performed on the data in this chapter. The chapter begins by describing quantitative measures which we computed using the dataset described in Chapter 3, then continues by discussing the results and their consequences for automatic tweet generation.

Our goal is to investigate what proportion of the text contained in the indicative tweets that we extracted can be found in the articles that they link to, in order to determine how well indicative tweet generation can be viewed as an extractive summarization problem. Table 4.1 gives an example of data where the tweet that was shared about the article does not come directly from the article text, while Table 4.2 shows a tweet that was almost entirely extracted from the text of the article, but changed slightly for the purpose of readability. Parts of it are extracted at the word or n-gram level.

Tweet	Are #Airlines doing enough with #Ebola? http://t.co/XExWwxmjnk #travel
Title	Could shortsighted airline refund policies lead to an outbreak?
Text	The deadly Ebola virus has arrived in the United States just in time for the holiday travel season, carrying fear and uncertainty with it...

Table 4.1: An example of a tweet, the title of the article it links to, and the text of the article, where the tweet cannot be extracted from the text.

Tweet	Officer Wilson will be returned to active duty if no indictment , says #Ferguson Police Chief http://t.co/zrRIBxMUYJ
Title	Jackson clarifies comments on Wilson's future status
Text	... Chief Jackson said if the grand jury does not indict Wilson , he will immediately return to active duty

Table 4.2: An example of a tweet, the title of the article it links to, and the text, where the tweet can be extracted from the text. The matched portions of the tweet and article are in bold.

We first compute the proportion of tweets that can be recovered directly from the article in its entirety (Section 4.1). Then, we calculate the degree of overlap in terms of unigrams and bigrams between the tweet and the text of the document (Sections 4.2, 4.3). We also compute the least common subsequences between the tweet and the document (Section 4.5).

In addition, we consider locality within the article when computing the overlap. For the unigram analysis, we performed a variant of the analysis, in which we computed the overlap within three-sentence windows in the source article (Section 4.4). We focused on locality in order to investigate whether sentence compression techniques could be applied to local context windows to generate the tweet.

These calculations are analogous to the ROUGE-1, -2 and -L style calculations that are standard in automatic evaluation of summarization systems. These results give an indication of the degree to which the tweet is extracted from the document text. Table 4.3 shows all the measures used by the analyses. They are ordered by the most restrictive measure, exact match, to the least restrictive measure, unigram match percentage. Unigram match in window and bigram match appear on the same level since they cannot easily be ordered with respect to each other.

For all of these analyses, stop words have been eliminated using NLTK's stop word list

Measure	Discussed in Section
Exact match	Section 4.1
LCS	Section 4.5
Bigram; Unigram match in window	Section 4.3; Section 4.4
Unigram	Section 4.2

Table 4.3: Measures used in analysis and corresponding sections where they are described, ordered from the most restrictive to least restrictive.

Stop Words
i, me, my, myself, we, our, ours, ourselves, yo, your, yours, yourself, yourselves, he, him, his, himself, she, her, hers, herself, it, its, itself, they, them, their, theirs, themselves, what, which, who, whom, this, that, these, those, am, is, are, was, were, be, been, being, have, has, had, having, do, does, did, doing, a, an, the, and, but, if, or, because, as, until, while, of, at, by, for, with, about, against, between, into, through, during, before, after, above, below, to, from, up, down, in, out, on, off, over, under, again, further, then, once, here, there, when, where, why, how, all, any, both, each, few, more, most, other, some, such, no, nor, not, only, own, same, so, than, too, very, s, t, can, will, just, don, should, now

Table 4.4: List of stop words in NLTK

from the tweet as well as the document, so that only the informative words are taken into consideration. The stop words are listed in Table 4.4. The comparisons were made without lemmatization or stemming, to adhere closely to existing work in extractive summarization, where the only modifications to the source text are removing discourse cue words or removing words by sentence compression techniques. The hashtags, references (@) and URLs from the tweets were all removed for the analyses.

4.1 Exact Match Calculations

We first checked for a complete substring match of the tweet in the text. This corresponds to the case where the tweet is in fact an extract taken from the linked article. Out of the 2471 unique instances of tweet and article pairs, a complete match was found only 23 times. In 9 cases out of these, the tweet text matched the title of the article, which our preprocessing tool did not correctly separate from the body of the article and introduced noise in the data. In the

Tweet	@PNHP: 6. Renounce punitive and counterproductive measures such as sealing the borders , http://t.co/LRLS2MhPRE #Ebola
Title	Physicians for a National Health Program
Text	As health professionals and trainees, we call on President Obama to take the following immediate steps to address the Ebola crisis... 6. Renounce punitive and counterproductive measures such as sealing the borders , and take steps to address the...

Table 4.5: Example where tweet is extracted as is from the text (matched portion in bold).

other cases, the text of the tweet appears in its entirety inside the body of the article. This suggests that the user chose to tweet the sentence that either seemed to be the most conclusive contribution of the article, or expressed the user’s opinion. An example of this is detailed in Table 4.5. The low number of exact matches found motivates us to check for partial match measures, specifically the n-gram measures that show how much of the tweet, if not whole, has been extracted from the article.

We also checked to see if the tweet text matched with the article titles that were separately extracted by the `newspaper` package. This was done in order to determine if tweets could be generated using the headline generation methods. We found that the tweet texts did not match with the titles in any of the remaining samples. However, the titles that could have been matched directly are accounted for in the 9 matched cases in the article text discussed earlier. Even though there are no exact matches, there might still be matches where the tweet is a slight modification of the headline of the article, and can be measured using a partial match measure. This difference could also have been caused by errors from the HTML to text extraction tool. This noise from the tool has been discussed in Section 3.2.

4.2 Percentage Match for Unigrams

Next, we computed the percentage match between the text of the tweet and the text in the article. This was a bag-of-words check using unigram overlap between the tweet and the document. Let $unigrams(x)$ be the set of unigrams for some text x , then u , the percentage

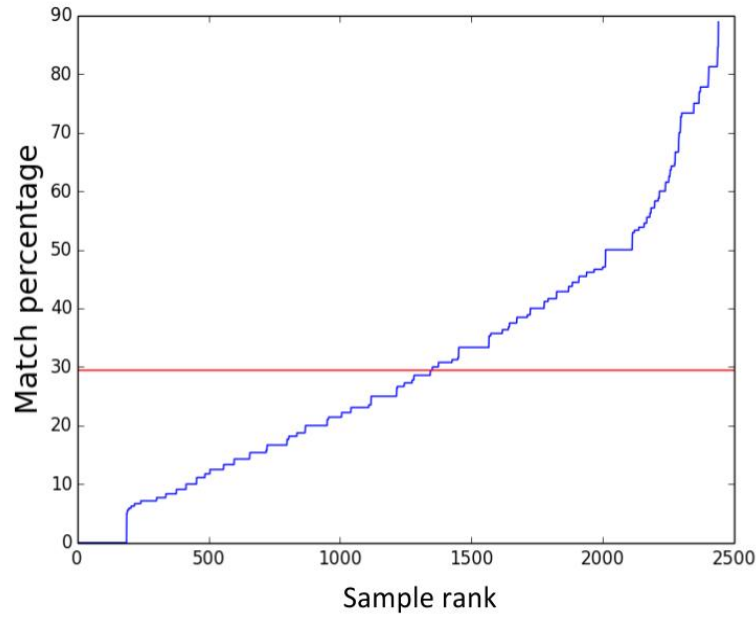


Figure 4.1: Distribution of unigram match percentage over unique tweet-article pairs, ordered from lowest percentage match to highest. The mean is 29.53%, indicated by the red horizontal line, with a standard deviation of 20.2%

of matching unigrams found between a given tweet, t and a given article, a , can be defined as

$$u = \frac{|unigrams(t) \cap unigrams(a)|}{|unigrams(t)|} * 100 \quad (4.1)$$

Figure 4.1 shows the distribution of percentage of unigram matches in the tweet and the article text with respect to each tweet and article pair. The mean match percentage is 29.53% and standard deviation is 20.2%. The mean of this distribution shows that the number of matched unigrams from a tweet in the article is fairly low. As an additional analysis, Figure 4.2 shows the number of articles with a certain number of matching unigrams. The graph shows that the most common number of unigrams matched was 2. The number of articles continues to decrease with higher unigrams matched. The slight rise at the end — more than 10 matched unigrams — is accounted for by the completely matched tweets described above.

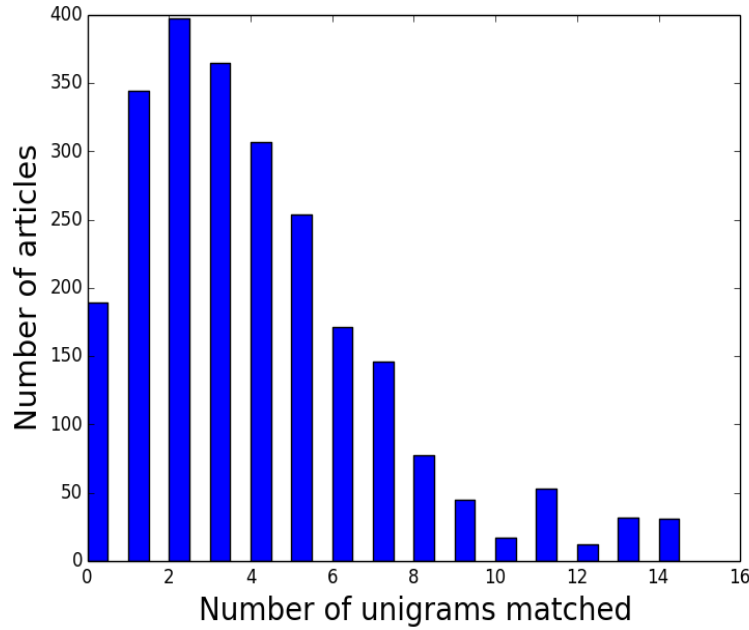


Figure 4.2: Histogram of number of unique tweet-article pairs vs number of unigrams matched. The mean number of unigrams matched per tweet-article pair is 3.9.

4.3 Percentage Match for Bigrams

Similar to the unigram matching techniques, the bigram percentage matching was also calculated. The text of the tweet was converted into bigrams and we then looked for those bigrams in the article text. The percentage was calculated similar to the unigram matching done earlier. For the set of bigrams for a text x , $bigrams(x)$, percentage of matching bigrams b for the tweet t and article a is:

$$b = \frac{|bigrams(t) \cap bigrams(a)|}{|bigrams(t)|} * 100 \quad (4.2)$$

Figure 4.3 shows the percentages of matched bigrams found. The mean is 10.73 with a standard deviation of 18.5. As seen in the figure, most of the tweet-article pairs have no matched bigrams. The percentage increase after this point is somewhat similar to that seen in the unigram match percentage section above.

Figure 4.4 shows the frequency of the number of tweet-article pairs for the number of

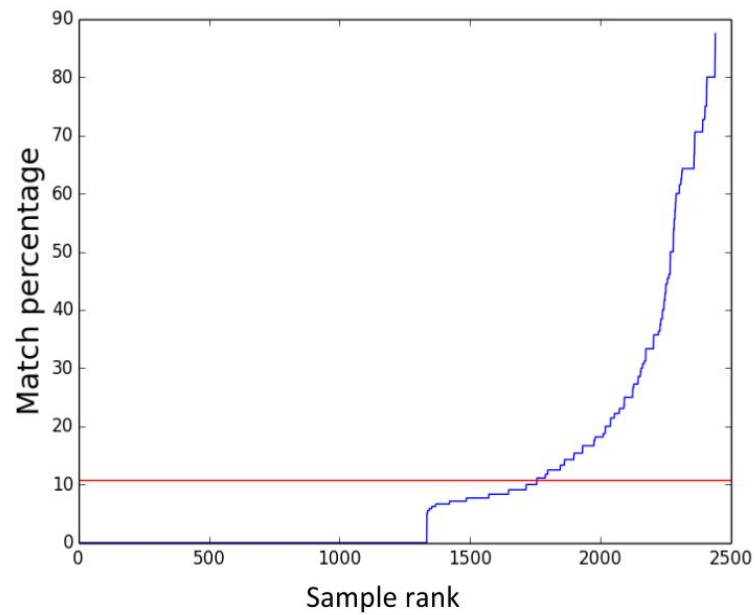


Figure 4.3: Distribution of bigram match percentage over the tweet-article pairs ordered from the lowest to the highest. The mean here is 10.73% shown by the red horizontal line, with a standard deviation of 18.5%

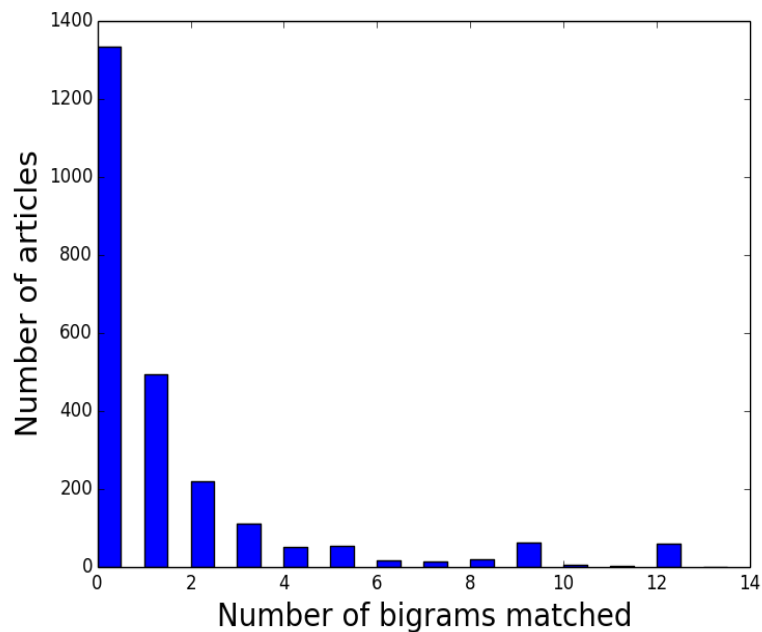


Figure 4.4: Histogram of number of unique tweet-article pairs vs number of bigrams matched. The mean number of bigrams matched per article is 1.9.

bigrams matched. There are no matched bigrams for most of the pairs. A smaller number of articles had one matched bigram, and the number decreased until the end, where it increases a little at more than 10 matched bigrams because of exact tweet matches.

The low percentage matches for unigrams and bigrams show that there are few common words between the tweet and the article. These low common n-gram percentages show that tweet generation cannot be approximated by extractive summarization well enough. We now perform some further analyses to confirm this observation.

4.4 Percentage Match Inside a Window in the Article Text

The next analysis checks for a significant word matching in a three-sentence window inside the article text. We used a three-sentence-long window using the sentence boundary information obtained during preprocessing. A window of three sentences was chosen to give a smaller context for the tweet to be extracted from than the entire article. The number was chosen as a moderate context window size; not too small to reduce it to the sentence level, and not too big for the context to be diluted. A window of five sentences was also experimented with, but there were no major differences in results between the three-sentence window and five-sentence window analysis. This analysis was performed to investigate whether a pseudo-extractive multi-sentence compression approach could convert a small number of sentences from the article into a tweet.

After the text of the window was extracted, we performed a similar analysis as the one in the unigram percentage matching, except on a smaller set of sentences. The matching percentages from all three-sentence windows in the articles were computed and the maximum out of these was taken for the final results. Let a sentence window, w_i , be the set of the words in three consecutive sentences starting from the sentence number i . For this window, the unigram match in the tweet t , and the window is the unigram match, u , calculated in Section 4.2. Then, the maximum match from all the windows, u^* is

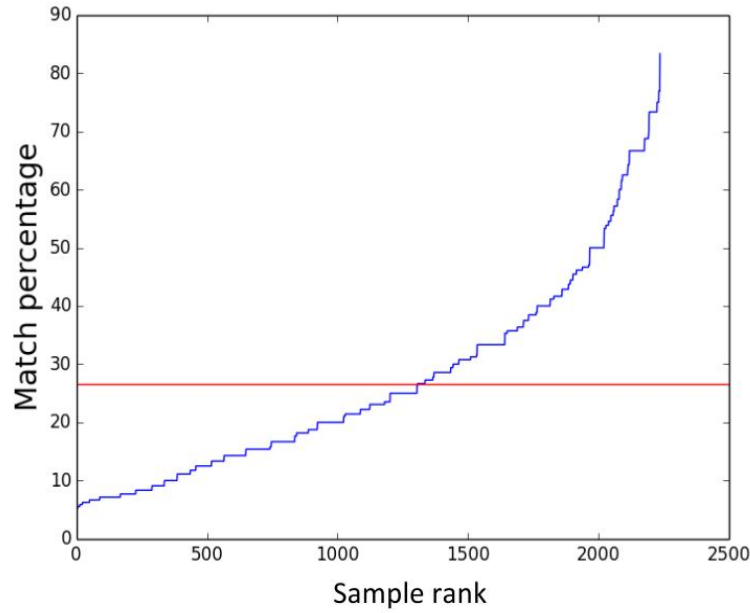


Figure 4.5: Percentages of common words in tweet and a three sentence window in the article. The maximum match from all percentages is chosen for an article. The red horizontal line is the mean is 26.6%, and standard deviation is 17%.

$$u^* = \max_{w_i \in S} u(t, w_i) \quad (4.3)$$

The result from this experiment is shown in Figure 4.5. Here, the mean of the values is 26.6% and standard deviation 17%. Again, this shows that only a small proportion of tweets can be generated even with an approach that combines unigrams from multiple sentences in the article.

If we look at the means of unigram matching for the entire document (29.5%) against that in a three sentence window (26.6%), there is only a difference of 2.9%. This difference translates to less than one unigram extracted from outside a small window in the article in the average case. These results seem to indicate that tweets can be extracted from a localized context almost as well as from the entire article. Nevertheless, we cannot use this information directly towards generating the tweet with sentence compression, since there is no easy way to determine where in the article the tweet has been extracted from. Also, even though the tweet can be extracted

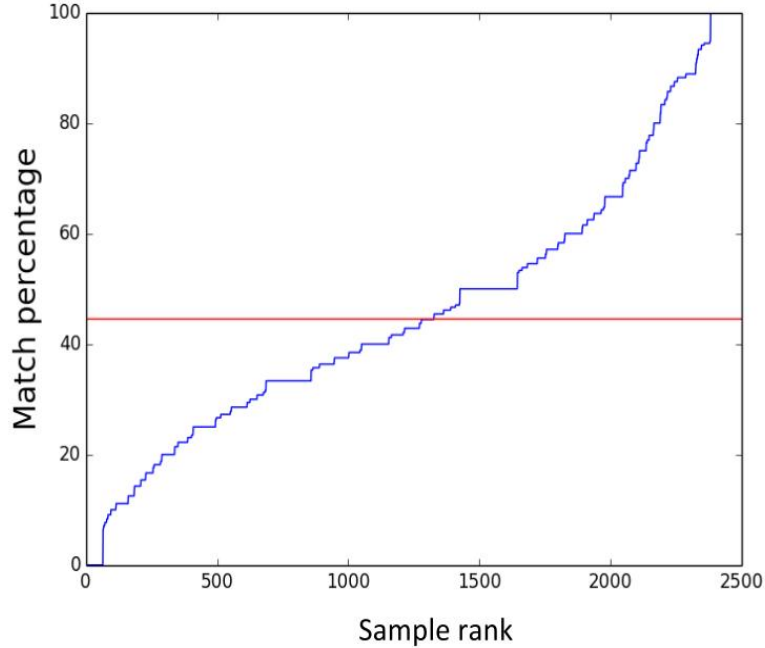


Figure 4.6: Percentages of words matching in tweet and document text using an LCS algorithm. Mean is 44.6%, which is shown by the red horizontal line, and standard deviation is 22.7%.

from a localized context as well as the whole article, the degree of extraction is still very low.

4.5 Longest Common Subsequence Match

The percentage match analyses were bag-of-words approaches that disregarded the order of the words inside the texts and tweets. To respect the order of the words in the sentence of the tweet, we also used the least common subsequence algorithm between the tweet text and the document text. This subsequence matching was done using the entire text of the article. The percentage match was calculated using the number of words in the tweet as the denominator.

If $lcs(t, a)$ is the longest common subsequence between the tweet t and article a , $len(x)$ is the length of the text x counted as number of words, then the percentage of match for the lcs as compared to the tweet, l is

$$l = \frac{len(lcs(t, a))}{len(t)} * 100 \quad (4.4)$$

These numbers are shown in Figure 4.6. The mean here is 44.6% and the standard deviation is 22.7%. The mean matching percentage in terms of length is higher than any of the earlier analyses. However, this number indicates that less than half the tweet is extracted from the article on an average. Note that the LCS matching percentage is higher than the unigram matching percentages because we consider the length of the matched string, instead of the set of matched words used in unigram matching. We discuss the implications of these results in the following section.

4.6 Interaction with Formality

As seen in the results of the analyses performed above, the tweets have little in common with the articles to which they are linked. This shows that extractive summarization algorithms can only recover a small proportion of the indicative tweets. One possible explanation for this result is that there may be a genre mismatch between the tweet and the linked article. Thus, while the semantic content may in some sense behave like an extractive summary, genre differences may result in low overlap scores. We thus computed the formality of the articles and investigated the impact of this score.

We assume that the formality of an article can be estimated by the formality of the words and phrases in the article. We used the formality lexicon of Brooke and Hirst (2013). 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 seed words appear in opposite halves of the graph, suggesting that we can use these seeds to determine if an article is formal or informal. The lexicon consists of words and phrases and their degree of formality. Thus, more formal words are marked on a positive scale and informal words like those occurring in colloquial language are marked on a negative scale.

Let the set of formality expressions from the lexicon be L , and the formality score for an

Lowest	Highest
#theforceawakens	#KevinVickers
#TaylorSwift	#erdogan
#winteriscoming	#apec

Table 4.6: Table of hashtags (broadly, topics) with highest and lowest formality according to the lexicon.

expression e be $score(e)$. Let the set of all substrings from the article $substrings(a)$ be S . Then, the formality score f for an article a is the number of formal expressions per 10 words in the article is

$$f = \frac{\sum_{e \in L \& e \in S} score(e)}{|unigrams(a)|} * 10 \quad (4.5)$$

The formality lexicon gave positive weights for formal expressions and negative for informal expressions. When we computed f using both formal and informal expressions, we found that the informal words predominated and “swamped” the signal of the formal words, leading to incomprehensible results. Thus, we discarded the informal words and used only the weights from the formal words in our final calculations. To check that these formality scores made sense intuitively, we calculated the average formality score for the articles belonging to each hashtag and ordered them, as shown in Table 4.6.

4.6.1 Examining Formality Scores with Respect to Match Percentages

The formality score for each article was correlated with the percentage matches obtained in our analyses in each case. All the correlation values were similar. Hence we only discuss the correlation value for longest common subsequence algorithm. The Pearson correlation value was 0.41, with a p-value of 7.08e-66, indicating that the interaction between formality and overlap was highly significant. Hence, we can conclude that the more formal the subject or the article, the better the tweet can be extracted from the article. Table 4.7 gives an example of the formality of the article, which has a low 4.2 formality words per 10 words, where the tweet is

Tweet	@globetoronto: Why Buffalo got clobbered with snow and Toronto did not. #weather #snowstorm http://t.co/gcwwDPZmX... http://t.co/BXY7EH6F3u ”
Title	What caused Buffalos massive snow and why Toronto got lucky
Text	Torontonians have long been the butt of jokes about calling in the army every time a few snow flurries whip by...

Table 4.7: Example of a tweet, title of the article where the formality of the article is lower, and the tweet is rephrased from the article.

not extracted from the article, but rephrased from the article instead.

These results confirm our hypothesis that formality, genre and source of text interact with the degree of extraction from the article, in a specific direction. We speculate that texts about formal events, or subjects such as politics would naturally have a serious tone, while tweets associated with less formal articles may contain more abbreviations and non-standard words or spellings, which decreases the amount of overlap. To counter this case, we tried experimenting with word normalization systems which could resolve words like ‘2mw’ or ‘4eva’ to ‘tomorrow’ and ‘forever’ respectively. Systems described by Yang and Eisenstein (2013) and Gouws et al. (2011) were tested with our data. Unfortunately, neither provided high enough performance to integrate with our analyses, and this remains something to reconsider upon the development of more accurate word normalization systems.

Chapter 5

User Study for Identifying Functions of Tweets

Chapter 4 described the analyses performed on the data and the resulting conclusion that only a small portion of indicative tweets can be recovered from the article they link to if viewed as an extractive summarization problem. This conclusion suggests that the next step should be to gain more knowledge about the relation between the tweet and the article, specifically why the user chose to write the tweet. This information would enable us to infer whether a tweet can actually be generated from summarizing the text in the article and if so, lead us to a strategy that would be the most appropriate for this task. We explore the task of collecting more information about the dataset in this chapter. We run a user study on Amazon Mechanical Turk, a crowdsourcing platform that enables researchers to put up huge samples of data along with surveys, quizzes, and simple tasks to be solved by people over the world. In this study, we asked the users whether a tweet is an advertisement or a summary of the article. We describe the process of running the user study and analysis of results obtained from the user study.

5.1 Functions of Tweets

One aspect of gaining more knowledge would be asking what the *‘function’* of the tweet is. While many possible definitions of ‘function’ are possible, our working definition of the term, introduced in Chapter 2, is why the user chose to share the article with the particular text written in the tweet. If the function of the tweet is known, it might be possible to predict the appropriateness of extractive methods for generating tweets. For example, if the function of the tweet is to be an advertisement for a particular product and the text contains the description of the features of the product, then intuitively, the tweet would likely be an extractive summary of the text. On the other hand, if the tweet were an attention grabbing title to merely bring traffic to the text, it would be more likely to be a generalized title, than a summary of the actual text. To identify these functions, we first examined multiple previous studies aiming to classify functions of tweets, discussed in Section 2.2.2. Based on these studies, we isolated the functions and finalized the questions to be asked about the tweets. The data was then given to human evaluators to annotate in a user study.

5.2 User Study Design

The following sections describe the details surrounding the design and execution of the user study.

5.2.1 Questions Used in the User Study

The questions to be asked in the user study with respect to each tweet-article pair came from our definition of the function of a tweet, described in Section 5.1. The final questions that were used are shown in Table 5.1.

The type of articles being referenced, and the possible reasons these might be shared gave the list of possible functions. Using these and earlier studies, we suggest a list of relevant functions of tweets for our data: promote a product or an article or convey information from

Question 1: Advertisement	Does the tweet explicitly encourage the reader to visit the link and read the original article?
Question 2: Summary	Does the tweet contain some information from the article, or summarize the article?

Table 5.1: Questions used in user study

the article. These functions will ideally help provide parameters for generating tweets. The idea behind these functions will be discussed in the following paragraphs.

Advertisement question If the tweet references a newspaper article, it might be promoting the article, in the sense of attracting people to read the article in detail. This kind of tweet would try to sensationalize the material. It could either tweet the headline of the article directly, or summarize the headline itself further, or simply say something to the effect of ‘Check this out’ or ‘This is worth a read’ and then further tag the article with the use of appropriate hashtags indicating the contents of the text.

Example: “Check out this article! {url}” or “Look what this says! {url}”

Summary question Secondly, the tweet could single out a particular piece of information or opinion directly from the article, either to agree with it or to express the importance of the sentence or phrase in the article according to the author of the tweet. It could also be a short summary of the text of the article either with the aim of inviting readers or just to inform readers about the contents of the article.

Example: “Winter approaching, ways to stay safe from flu season: {url}”

These questions were presented separately in two different studies. We found a possibility that workers were viewing the questions in the pilot studies as either/or questions, where the answer to only one of them could be ‘yes’. Thus, separating the questions guaranteed an unbiased opinion about each question without any assumptions.

A third possible question of whether the tweet expressed an emotion towards the article, and if so, whether it was a positive or negative emotion was also considered. However, it was

not included in the final study, and will be discussed in Section 5.2.4

5.2.2 Running the User Study

Figure 5.1 and Figure 5.2 show an example of what a HIT (task on Mechanical Turk) for the two questions looked like to the user, respectively. For each of the questions, the first part of the figure shows the instructions involved as well as examples for every possible answer to the question asked. The tweet, the title that was extracted and the entire text of the article was then presented to help the workers make a decision about their answer.

5.2.3 Qualification Details

Mechanical Turk assigns qualifications to workers based on their skill level in answering questions and the percentage of accepted answers. Workers with superior skills and higher accuracy while answering HITs are given the qualification ‘Master’ by Mechanical Turk, which was the qualification used for these pilot studies.

Three separate opinions from the workers were gathered for each tweet and article pair and the inter-annotator agreement was calculated using Fleiss’ kappa measure of agreement (Geertzen, 2012) between the raters.

We found that the way to obtain the best quality results was to add an additional level of qualification over the default Master’s qualification. The qualification test contained three tweet-article pairs from the dataset that had an obvious category for function of tweets. The test was conducted separately for the two questions asked. These tests were presented to the workers, and only workers with a 100% score were allowed to work on the data. These tests ensured that the worker had a complete understanding of the task at hand. A separate pilot study with these qualifications and questions structure gave promising results, and the user study was then run on the entire dataset.

Instructions
<p>These questions relate to the function of the tweet with respect to the article. Each HIT contains a tweet, an article and the title of the article. Answer this question about the tweet and the article based on the contents of both :</p> <p>Does the tweet explicitly encourage the reader to visit the link and read the original article? It might directly refer to the article itself, rather than the contents of the article. Note that this includes advertisements for products, as well as just plain links or hashtags. Another case is if it is exactly the same as the title of the article. For example,:</p> <p>Yes</p> <ul style="list-style-type: none"> <p>Tweet : Adidas Gazelle VIVID PINK WHITE EXCLUSIVE Trainers VH1 #adidas #fashion #TheForceAwakens http://t.co/LhEYVi4oi8 http://t.co/sO1pi5kBe</p> <p>Title : Adidas Gazelle Og Mid VIVID PINK WHITE EXCLUSIVE Trainers Shoes</p> <p>Article : First issued in 1968 as an all around training shoe the Gazelle had a streamlined profile and innovative cushioning that made it a hit among top athletes. This reissue...</p> <p>Reason -> Attention grabbing title with name of the product</p> <p>Tweet : Why I love #montythepenguin http://t.co/shh8XfxVlp @UKBlog_RT @FemaleBloggerRT #careers #marketing</p> <p>Title : The John Lewis Effect</p> <p>Article : Last week John Lewis blessed us all with their latest Christmas advert. This year's ad features Monty the Penguin or should I say...</p> <p>Reason -> Tweet gives a positive review of the content and encourages reading of the article</p> <p>No</p> <ul style="list-style-type: none"> <p>Tweet : RT @ServcorpUSA: This story warms our heart. #RealEstate Broker Proposes at Servcorp office on #1WTC's 85th floor. http://t.co/7boH8NwUOL</p> <p>Title : Man proposes at 1 World Trade Center's 85th floor</p> <p>Article : These lovers are on top of the world.A Manhattan real estate broker proposed to his girlfriend of five years on 1 World Trade Centers 85th floor Tuesday and she said yes.Jeffrey Carlson...</p> <p>Reason -> This is a no because it is more about the emotions felt because of the content and the summary of what happened.</p> <p>Tweet : RT @csmonitor: #SCOTUS will consider whether the IRS overstepped by awarding tax credits under #Obamacare: http://t.co/L4JWyn0RS @WarrenRi</p> <p>Title : Supreme Court agrees to hear new challenge to Obamacare (+video)</p> <p>Article : The Supreme Court agreed Friday to consider whether the IRS overstepped its authority when it permitted the agency to award tax credits to people who signed...</p> <p>Reason -> Again, a clear summary of what's in the article.</p> <p>Link to our ethics consent form. (https://www.dropbox.com/s/00uvcoe58wb3wbr/Consentform.pdf?dl=0)</p>

Tweet

Hamas has effectively established a command post in Turkey #pvv #feiten #Erdogan <http://t.co/itwzLIX5wE>

Article Title

Israel nabs West Bank terror network 'commanded from Turkish soil'

Figure 5.1: (a) The first question (Advertisement) posed for each sample asked to the users.

Article

Israel has arrested dozens of members of a Hamas terror network operating throughout the West Bank in recent weeks who were planning a series of attacks against Israeli targets senior Palestinian officials told The Times of Israel. The network they said was funded and directed by Hamas officials in Turkey who have set up a de facto command center in the Muslim country. The network was similar in its operational characteristics to one uncovered in August during the war with Hamas in the Gaza Strip the officials said Thursday night adding that according to information received from Israel this terror ring was even larger. Its operatives had already attempted several attacks against Israel they added but they had all failed. As with the previous network the man behind the terrorist grouping was Saleh al-Arouri a Hamas leader who was deported from the West Bank to Turkey in 2010 the sources said. Arouri they said built up and funded the network and has effectively established a Hamas command post in Turkey which is leading terror efforts in the West Bank. Arouri is reportedly aided by dozens of operatives some of whom were deported

1. Does the tweet explicitly encourage the reader to visit the link and read the original article?

- ☐ Yes
☐ No

Comments (Optional)

If you answered "No" to question 1, please comment what you think the tweet does in one or two words:

Figure 5.1: (b) The first question (Advertisement) posed for each sample asked to the users.

Instructions
<p>These questions relate to the function of the tweet with respect to the article. Each HIT contains a tweet, an article and the title of the article. Answer a this question about the tweet and the article based on the contents of both :</p> <p>Does the tweet contain some information from the article, or summarize the article?</p> <p>Examples :</p> <p>Yes</p> <ul style="list-style-type: none"> <p>Tweet : So Wrong - #Putin and #Obama break at the #G20 to cuddle some koalas http://t.co/xfTXt3Xj0P via @mashable</p> <p>Title : Putin and Obama break at the G20 to cuddle some koalas</p> <p>Article : It looks like the G20 Leaders' Summit where heads of state come together to discuss pressing international issues isn't all business. Apparently ...</p> <p>Reason -> Has parts of the article text and title, which summarizes the event described in the article.</p> <p>Tweet : Can Philae get a better 'kick' this time? Comet #67P's vapor-jets may toss #Philae Lander to a better location http://t.co/jZ0wnDeAhp</p> <p>Title : Adventures of European comet lander may not be over</p> <p>Article : Gas jets from inside a comet hosting Europe's Philae lander may launch the hibernating probe out of its ditch and back into sunlight for a battery recharge a former mission manager said on Monday. The European Space Agency...</p> <p>Reason -> The tweet is again summarizing what is described in more detail in the article, even though it is not the same as the article.</p> <p>No</p> <ul style="list-style-type: none"> <p>Tweet : Why I love #montythepenguin http://t.co/shh8XfxVlp @UKBlog_RT @FemaleBloggerRT #careers #marketing</p> <p>Title : The John Lewis Effect</p> <p>Article : Last week John Lewis blessed us all with their latest Christmas advert. This year's ad features Monty the Penguin or should I say ...</p> <p>Reason -> Tweet only has a positive review of the content of the article.</p> <p>Tweet : Adidas Gazelle VIVID PINK WHITE EXCLUSIVE Trainers VH1 #adidas #fashion #TheForceAwakens http://t.co/LhEYVi4oi8 http://t.co/sO1ipi5kBe</p> <p>Title : Adidas Gazelle Og Mid VIVID PINK WHITE EXCLUSIVE Trainers Shoes</p> <p>Article : First issued in 1968 as an all around training shoe the Gazelle had a streamlined profile and innovative cushioning that made it a hit among top athletes. This reissue...</p> <p>Reason -> Attention grabbing title that announces the product.</p> <p>Link to our ethics consent form. (https://www.dropbox.com/s/00uvcoe58wb3wbr/Consentform.pdf?dl=0)</p>

Tweet

RT @CIAwesome: It's time to speak out against sexual assault <http://t.co/4gQRIZTMrp> #yyc #support #consent #beenrapedneverreported

Article Title

Its time to speak out against sexual assault

Figure 5.2: (a) The second question (Summary) asked for each sample.

Article

Never before has the subject of sexual assault been such a national focus in Canada. The allegations against Jian Ghomeshi have brought the topics of consent and victim blaming to the forefront of what is a long-overdue discussion. For victims it has also been a difficult couple of weeks with both traditional and social media acting as constant reminders of how often victims are blamed when they choose to speak up. However the way we talk about sexual assault is beginning to change the more we talk about it the more people will be educated about the problem and the sooner we can all begin to work towards positive change. When I first heard the whispers about the allegations against Jian Ghomeshi in late October I was saddened but not shocked according to Statistics Canada over 500000 sexual assaults are reported in this country every year. They also note that 91 per cent of sexual assaults are not reported. The majority of women who experience harassment assault and rape are women in my demographic (post-secondary students aged 18-24). Personal experience and testimonials from friends confirm that sexual violence

1. Does the tweet contain some information from the article, or summarize the article?

- ☐ Yes
☐ No

Comments (Optional)

If you answered "No" to question 1, please comment what you think the tweet does in one or two words:

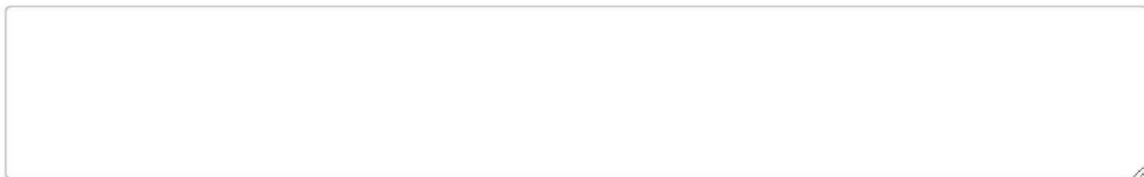


Figure 5.2: (b) The second question (Summary) asked for each sample.

5.2.4 Pilot Studies

Before finalizing the design of the study, we considered various different options, which will be discussed further in this section.

Attempts at tagging data Our first attempt at tagging the dataset was made using the following scheme: a sample set of 100 articles were tagged by two people, based on the title, the article and the tweet. The tags used in the preliminary tagging were ‘evaluative’, ‘descriptive’, and ‘mixed’. ‘Evaluative’ text is a more opinionated text, while ‘Descriptive’ text is non-evaluative, containing, for example, a narration of an event or an explanation about a certain object or event. A ‘mixed’ category was also added during tagging to accommodate some articles that could fall in either category. It was observed that this classification was rather subjective. Liu and Zhang (2012) give a detailed analysis on classifying evaluative vs descriptive texts, which is called as subjectivity classification. They point out that it is a bad idea to classify on a sentence level in complex sentences, since a sentence and by extension a text might be factual as well as evaluative, which was the original problem while tagging. This tagging scheme did not yield much success in terms of gleaning information over the entire dataset.

Pilot studies for the user study Four pilot studies were conducted sequentially on 100 different randomly selected tweet and article pairs each time, using each to improve the design for the next study. For these pilot studies, we tried various combinations of questions and qualifications, which are a type of grading system based on experience and ability for workers on Mechanical Turk.

A third possible function of tweets Another possible question that was briefly mentioned in Section 5.2.1 is whether the tweet expresses an opinion about something in the article. It could be a positive or a negative comment about the contents of the article, or an agreement or disagreement about the events, thoughts, and opinions expressed in the article.

Example: “This is a terrible analysis. {url}” or “Listening to this album on repeat! {url}”

However, this question was dropped going forward with the studies because of the difficulty in distinguishing the difference between emotion towards the article vs emotion reiterated from the article itself.

5.3 Results and Analysis for the User Study

This section describes the analysis of results obtained from the user study.

The total number of ‘yes’ votes for each of the questions are shown in Table 5.2. Table 5.3, Table 5.4, Table 5.5 and Table 5.6 all show examples where for one of each question, all three workers agreed on their answers to each of the two questions. The Fleiss’ kappa for the first study, showing the indicativeness of the tweet was 0.147 and the kappa for the second study, showing the informativeness of the tweet, was 0.208.

We then analyzed the answers obtained from the study by correlating them with the overlap measures defined in Chapter 4 with the help of the Mann Whitney U test (Mann and Whitney, 1947; Wilcoxon, 1947). The Mann Whitney U test considers two groups of ratings and then analyses them in terms of rankings to infer how they corroborate. Each of the following tables shows a result for the Mann Whitney U test for a study pertaining to one of the questions, and a corresponding analysis: unigram, bigram, or LCS matching percentages. Table 5.7 and Table 5.8 show the test results for unigram match percentages, Table 5.9 and Table 5.10 for bigram match percentages, and Table 5.11 and Table 5.12 for longest common subsequence match percentages.

Mechanical Turk presents each question for each tweet-article pair to three different workers. Thus, every tweet-article pair has three different opinions for each question asked. We split all the tweet-article pairs from the data into two different groups. We form the groups for the Mann-Whitney U test using the above information. The first split considers zero ‘yes’ votes in the answers as one group and three ‘yes’ votes as another group. The second split considers

Questions	0 ‘yes’ votes	1 ‘yes’ vote	2 ‘yes’ votes	3 ‘yes’ votes
Q1: Advertisement	53	340	942	1068
Q2: Summary	29	176	703	1495

Table 5.2: Analysis of user study results

Tweet	RT @WSJ: In #CometLanding Philae probe bounced and settled in area that could hinder its research. http://t.co/6lfg3p9XG1 http://t.co/A6fi
Title	Rosetta Mission Probe Landed on Comet in Shadow of Cliff
Text	The historic Philae comet probe hit its target but then unexpectedly bounced twice settling in the shadow of a cliff that could hinder its research new images sent back Thursday showed.Philae is designed to run a suite...

Table 5.3: Example where all three workers said it was not an advertisement.

Tweet	#GalaxyNote3 #Lollipop - SamMobile has been teasing us with a number of unfinished builds for a few http://t.co/A0IKYsk4g3 #Samsung
Title	Samsung GALAXY Note 3’s Android Lollipop Update Surfaces
Text	SamMobile has been teasing us with a number of unfinished builds for a few months now. This indicates...

Table 5.4: Example where all three workers said the tweet was an advertisement for article.

Tweet	”RT @jakbarali: So my partner Gillian Hnatiw and I had something to say about #VAW #LoriDouglas and #Ghomeshi. http://t.co/6X2zMtCAM0
Title	”Victim-blaming couched as legitimate judicial inquiry”
Text	Ghomeshi himself broke the first wave of the story when he took to Facebook to decry the CBCs decision to terminate him...

Table 5.5: Example where all three raters said the tweet was not a summary.

Tweet	RT @PopCulturPriest: Doing a story on California’s lottery for @americamag I discovered #JohnOliver’s story had some troubling errors: http://t.co/6X2zMtCAM0
Title	Blowing The Dismount: Last Week Tonight Fudges Its Lottery Story
Text	Sunday night on the season finale of HBOs new news show Last Week Tonight anchor John Oliver spent half the show...

Table 5.6: Example where all three raters agreed the tweet was a summary.

zero or one ‘yes’ votes out of three in one group and two or three ‘yes’ votes out of three in the other group. For each of these studies for each sample set configuration, the U statistic and the p value are shown. The final two columns in both tables show the mean of values in each of the groups used in the test.

Groups considered	U statistic	p value	Mean of values for Group 1	Number of samples in Group 1	Mean of values for Group 2	Number of samples in Group 2
Group 1: 0 'yes' votes Group 2: 3 'yes' votes	28104	0.931413	27.44	53	28.21	1068
Group 1: 0 or 1 'yes' votes Group 2: 2 or 3 'yes' votes	406355.5	0.365219	30.65	393	29.23	2010

Table 5.7: Mann Whitney U test results for the advertisement question(indicativeness): Unigram Match

Groups considered	U statistic	p value	Mean of values for Group 1	Number of samples in Group 1	Mean of values for Group 2	Number of samples in Group 2
Group 1: 0 'yes' votes Group 2: 3 'yes' votes	12211	0.000055	16.69	29	31.07	1495
Group 1: 0 or 1 'yes' votes Group 2: 2 or 3 'yes' votes	193411	0.000791	25.08	205	29.87	2198

Table 5.8: Mann Whitney U test results for the summary question(informativeness): Unigram Match

Groups considered	U statistic	p value	Mean of values for Group 1	Number of samples in Group 1	Mean of values for Group 2	Number of samples in Group 2
Group 1: 0 'yes' votes Group 2: 3 'yes' votes	27871	0.851388	8.31	53	9.21	1068
Group 1: 0 or 1 'yes' votes Group 2: 2 or 3 'yes' votes	406313	0.378553	12.19	393	10.29	2009

Table 5.9: Mann Whitney U test results for the advertisement question(indicativeness): Bigram Match

Groups considered	U statistic	p value	Mean of values for Group 1	Number of samples in Group 1	Mean of values for Group 2	Number of samples in Group 2
Group 1: 0 'yes' votes Group 2: 3 'yes' votes	15006	0.004541	3.88	29	11.46	1494
Group 1: 0 or 1 'yes' votes Group 2: 2 or 3 'yes' votes	201755.5	0.013592	8.07	205	10.84	2197

Table 5.10: Mann Whitney U test results for the summary question(informativeness): Bigram Match

Groups considered	U statistic	p value	Mean of values for Group 1	Number of samples in Group 1	Mean of values for Group 2	Number of samples in Group 2
Group 1: 0 'yes' votes Group 2: 3 'yes' votes	26440.5	0.418424	42.16	53	44.24	1068
Group 1: 0 or 1 'yes' votes Group 2: 2 or 3 'yes' votes	392910	0.870236	44.66	393	44.69	2010

Table 5.11: Mann Whitney U test results for the advertisement question(indicativeness): Longest Common Subsequence

Groups considered	U statistic	p value	Mean of values for Group 1	Number of samples in Group 1	Mean of values for Group 2	Number of samples in Group 2
Group 1: 0 'yes' votes Group 2: 3 'yes' votes	18466	0.171255	38.4	29	44.47	1495
Group 1: 0 or 1 'yes' votes Group 2: 2 or 3 'yes' votes	217196.5	0.393999	43.16	205	44.83	2198

Table 5.12: Mann Whitney U test results for the summary question(informativeness): Longest Common Subsequence

The p-values for Table 5.7, Table 5.9 and Table 5.11 show non-significant results for both sets of groups for the first question, the indicativeness of the tweet. The U statistic for each case is very high and the results show a $p > 0.05$. We thus fail to reject the null hypothesis that the two sets were pulled from the same distribution. For all these cases, the means of the two groups are very close to the means for the respective analysis, and to each other. Mean for Unigram match is 29.53%, mean for bigram match is 10.73% and the mean for LCS match is 44.6% as seen in Chapter 4.

The p-values for unigram and bigram match for the second question, indicating the informativeness of the summary, shown in Table 5.8 and Table 5.10 are both significant, with $p < 0.05$, especially so for the first arrangement of groups where group 1 is zero ‘yes’ votes and group 2 is three ‘yes’ votes. Based on the result of the p-values, we can conclude that these samples are drawn from different populations. If we look at the means of the values in each case, they are sufficiently different, with the mean of the first group being significantly smaller than the mean of the values in the second group. Table 5.12 also shows a slight difference in the means when zero vs three ‘yes’ votes were considered as the sample set configuration. The U-statistic and p-value are both the least in this case for longest common subsequence results. However, no significant result can be drawn from this since the p-value is still quite high. It is possible that the non-significant result can be explained by the fact that the LCS is a lot more flexible for accommodating words from the overall article, and thus while the means of the two groups show difference in the right direction, the p-value is still too high to conclude anything significant.

5.3.1 Conclusions from the User Study

The significant results from Table 5.8 and Table 5.10 represent evidence that tweets that are informative and tweets that are not informative have different levels of extractiveness from their source article. However, the evidence does not support the fact that whether a tweet is an advertisement interacts with the extractiveness of the tweet. Further studies would be required

to come to a conclusion about this type of summary classification based on function, and how it interacts with extractiveness of the summary. The study shows a promising direction for further studies on the function of tweets.

An important outcome of this chapter is the generation of a human-tagged dataset of tweet and article pairs, based on the indicativeness and informativeness of the tweets with respect to the article text.

The question of whether a tweet summarizes the content of the article gave mostly positive answers, suggesting that according to the workers, if the tweet contained a link to article, it was an indicative summary in most cases. However, according to the extractiveness calculated earlier in Chapter 4, the tweets were not extracted from the articles to a large extent. With the results from the user study performed in this chapter, we can see that even when the tweet is used informatively, extractive methods have an upper bound that is still low, similar to what was obtained earlier. This reinforces the earlier conclusion of a need for a more sophisticated tool that summarizes the contents of the article for tweet generation.

Chapter 6

Conclusion

We have described a study that investigates whether indicative tweet generation can be viewed as an extractive summarization problem. By analyzing a dataset of indicative tweets that we collected using measures inspired by extractive summarization evaluation, we find that most tweets cannot be recovered from the article that they link to, demonstrating a limit to the effectiveness of extractive methods.

We further performed an analysis to determine the role of formality differences between the source article and the Twitter genre. We find evidence that formality is an important factor, as the less formal the source article is, the less extractive the tweets seem to be. Future methods that can change the level of formality of a piece of text without changing the contents will be needed, as will those that explicitly consider the intended use of the tweet.

Finally, we conducted a study to determine whether the function of the tweet towards the article was a factor in the degree to which the tweet was extracted from the article. The analyses performed in Chapter 4 show that a small percentage of tweets can be extracted from articles. The user study further confirms that a majority of articles are summaries of the articles, according to the workers. This shows that it is worth pursuing abstractive summarization as a way to generate tweets. We have consequently generated a dataset of tweets and articles categorized by topic, and asked users to tag them according to whether the tweet is an ad-

vertisement encouraging the user to click on and read the entire article, or a summary of the article. This generated dataset of tagged tweets and articles is an important contribution of the thesis, and can be used in further studies towards identifying functions of tweets and also in tweet generation.

6.1 Future Work

6.1.1 Study Functions and Intents

Our studies of communicative functions have explored two aspects of the functions of tweets. It would be worthwhile to further explore the reasons for writing tweets, to be able to classify them, and use this information further as parameters for advertisements or personalized feeds. Analysis of the text and the tweet itself in conjunction with the various intents described in Sinclair and Ball (1996) would help to solve the problem.

6.1.2 A Structure for Generating Tweets

The final goal would be the ability to generate a tweet based on the text of the article or a blog, possibly with the help of a parameter: a communicative goal mentioned above. The communicative goal would help establish the context in which the tweet would be used and therefore the kind of tweet that needs to be generated from the text.

6.1.3 Parameterized summarization

A broader parameterized text summarization system would be an excellent generalization of the tweet generation process. This would not only include a way to generate a summary according to the way in which the summary would be used, but also consider what the summary intends to convey from the text. For example, a summary could be converted to a higher or a lower level of formality for publishing to different outlets. A summary posted on a social media

platform would be less formal whereas a summary posted on a blog would be comparatively more formal.

Bibliography

- Banerjee, N., Chakraborty, D., Joshi, A., Mittal, S., Rai, A., and Ravindran, B. (2012). Towards analyzing micro-blogs for detection and classification of real-time intentions. In *Proceedings of International Conference on Web and Social Media*, pages 391–394.
- Brooke, J. and Hirst, G. (2013). A multi-dimensional bayesian approach to lexical style. In *Human Language Technologies-North American Chapter of the ACL*, pages 673–679.
- Chakrabarti, D. and Punera, K. (2011). Event summarization using tweets. In *Proceedings of International Conference on Web and Social Media*, volume 11, pages 66–73.
- Cheung, J. C. K. and Penn, G. (2013). Towards robust abstractive multi-document summarization: A caseframe analysis of centrality and domain. In *Association for Computational Linguistics*, pages 1233–1242.
- Chu, Z., Gianvecchio, S., Wang, H., and Jajodia, S. (2012). Detecting automation of twitter accounts: Are you a human, bot, or cyborg? *IEEE Transactions on Dependable and Secure Computing*, 9(6):811–824.
- Conroy, J. M., Schlesinger, J. D., and O’Leary, D. P. (2006). Topic-focused multi-document summarization using an approximate oracle score. In *Proceedings of the COLING/ACL on Main Conference Poster Sessions*, pages 152–159. Association for Computational Linguistics.

- Geertzen, J. (2012). Inter-rater agreement with multiple raters and variables. <https://nlp-ml.io/jg/software/ira/>.
- Ghosh, R., Surachawala, T., and Lerman, K. (2011). Entropy-based classification of 'retweeting' activity on twitter. *Social Network Mining and Analysis*.
- Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., and Smith, N. A. (2011). Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 42–47. Association for Computational Linguistics.
- Gómez-Adorno, H., Pinto, D., Montes, M., Sidorov, G., and Alfaro, R. (2014). Content and style features for automatic detection of users intentions in tweets. In *Advances in Artificial Intelligence-IBERAMIA 2014*, pages 120–128. Springer.
- Gouws, S., Metzler, D., Cai, C., and Hovy, E. (2011). Contextual bearing on linguistic variation in social media. In *Proceedings of the Workshop on Languages in Social Media*, pages 20–29. Association for Computational Linguistics.
- Hahn, U. and Mani, I. (2000). The challenges of automatic summarization. *Computer*, 33(11):29–36.
- Han, B. and Baldwin, T. (2011). Lexical normalisation of short text messages: Makn sens a# twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 368–378. Association for Computational Linguistics.
- He, L., Sanocki, E., Gupta, A., and Grudin, J. (2000). Comparing presentation summaries: Slides vs. reading vs. listening. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 177–184. ACM.

- Inouye, D. and Kalita, J. K. (2011). Comparing twitter summarization algorithms for multiple post summaries. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom)*, pages 298–306.
- Kaufmann, M. and Kalita, J. (2010). Syntactic normalization of twitter messages. In *International conference on natural language processing, Kharagpur, India*.
- Knight, K. and Marcu, D. (2002). Summarization beyond sentence extraction: A probabilistic approach to sentence compression. *Artificial Intelligence*, 139(1):91–107.
- Kong, L., Schneider, N., Swayamdipta, S., Bhatia, A., Dyer, C., and Smith, N. A. (2014). A dependency parser for tweets. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1001–1012, Doha, Qatar. Association for Computational Linguistics.
- Kouloumpis, E., Wilson, T., and Moore, J. D. (2011). Twitter sentiment analysis: The good the bad and the omg! *Proceedings of International Conference on Web and Social Media*, 11:538–541.
- Lin, C.-Y. (2004a). Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81.
- Lin, C.-Y. (2004b). Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81.
- Lin, C.-Y. and Och, F. (2004). Looking for a few good metrics: Rouge and its evaluation. In *NTCIR Workshop*.
- Liu, B. and Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In *Mining Text Data*, pages 415–463. Springer.
- Liu, F. and Liu, Y. (2009). From extractive to abstractive meeting summaries: Can it be done by

- sentence compression? In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 261–264. Association for Computational Linguistics.
- Lloret, E. and Palomar, M. (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.
- Lofi, C. and Krestel, R. (2012). iparticipate: Automatic tweet generation from local government data. In *Database Systems for Advanced Applications*, pages 295–298. Springer.
- Mani, I. (2001). *Automatic Summarization*, volume 3. John Benjamins Publishing.
- Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, pages 50–60.
- Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (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.
- Mohammad, S. M., Kiritchenko, S., and Martin, J. (2013a). Identifying purpose behind electoral tweets. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, pages 1–9. ACM.
- Mohammad, S. M., Kiritchenko, S., and Zhu, X. (2013b). Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. *arXiv preprint arXiv:1308.6242*.
- Nenkova, A. and McKeown, K. (2012). A survey of text summarization techniques. In *Mining Text Data*, pages 43–76. Springer.
- Nichols, J., Mahmud, J., and Drews, C. (2012). Summarizing sporting events using twitter. In *Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces*, pages 189–198. ACM.

- O'Connor, B., Krieger, M., and Ahn, D. (2010). Tweetmotif: Exploratory search and topic summarization for Twitter. In *Proceedings of International Conference on Web and Social Media*, pages 384–385.
- Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., Schneider, N., and Smith, N. A. (2013). Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–390, Atlanta, Georgia. Association for Computational Linguistics.
- Ritter, A., Clark, S., Etzioni, O., et al. (2011). Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534. Association for Computational Linguistics.
- Sidhaye, P. and Cheung, J. C. K. (2015). Indicative tweet generation: An extractive summarization problem? In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, pages 138–147.
- Sinclair, J. and Ball, J. (1996). Preliminary recommendations on text typology. *EAGLES (Expert Advisory Group on Language Engineering Standards)*.
- Wang, J., Cong, G., Zhao, X. W., and Li, X. (2015). Mining user intents in twitter: A semi-supervised approach to inferring intent categories for tweets. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 339–345.
- Wang, L., Cardie, C., and Marchetti, G. (2014). Socially-informed timeline generation for complex events. In *Proceedings of Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*, pages 1055–1065.
- Wei, Z. and Gao, W. (2014). Utilizing microblogs for automatic news highlights extraction. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 872–883.

- Wilcoxon, F. (1947). Probability tables for individual comparisons by ranking methods. *Biometrics*, 3(3):119–122.
- Yang, Y. and Eisenstein, J. (2013). A log-linear model for unsupervised text normalization. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, pages 61–72.