Comparing Algorithms for Microblog Summarisation

Stuart Mackie[1], Richard McCreadie[2], Craig Macdonald[2], and Iadh Ounis[2]

School of Computing Science, University of Glasgow, G12 8QQ, UK
[1]s.mackie.1@research.gla.ac.uk, [2]{firstname.lastname}@glasgow.ac.uk

Abstract. Event detection and tracking using social media and user-generated content has received a lot of attention from the research community in recent years, since such sources can purportedly provide up-to-date information about events as they evolve, e.g. earthquakes. Concisely reporting (summarising) events for users/emergency services using information obtained from social media sources like Twitter is not a solved problem. Current systems either directly apply, or build upon, classical summarisation approaches previously shown to be effective within the newswire domain. However, to-date, research into how well these approaches generalise from the newswire to the microblog domain is limited. Hence, in this paper, we compare the performance of eleven summarisation approaches using four microblog summarisation datasets, with the aim of determining which are the most effective and therefore should be used as baselines in future research. Our results indicate that the SumBasic algorithm and Centroid-based summarisation with redundancy reduction are the most effective approaches, across the four datasets and five automatic summarisation evaluation measures tested.

1 Introduction

Microblogging services (e.g. Twitter[1]) provide a platform for people and organisations to share up-to-date information about many topics, particularly news and current events. Such social media services are facilitating a shift towards real-time news reporting and discussion of events by the public and organisations. As a result, end-users and journalists leverage social media to monitor and track events as they evolve over time [2, 14]. However, due to the high volume and velocity of messages posted to social media streams[2], there may be vastly more posts published than users could ever read. This means that users would find it very difficult to keep up-to-date with events of interest.

To tackle this problem, summarisation algorithms have been proposed such as SumBasic [10] or Hybrid TF-IDF [12]. Automatic text summarisation techniques [9, 13] must algorithmically decide what is the essential information from the input text(s) that should be reported to the user as a summary. However, to-date, there has been little research regarding how different summarisation algorithms compare in terms of absolute performance, for the task of microblog summarisation. The only recent comparison of summarisation algorithms for microblog summarisation was performed by Sharifi *et al.* [12]. This study indicated that relatively simple term-frequency algorithms, such as Hybrid-TFIDF, offered reasonable summary effectiveness – but was limited in scope,

[1] http://twitter.com/
[2] https://blog.twitter.com/2014/celebrating-sb48-on-twitter

		Components			
Category	**Approach**	**Representation**	**Scoring**	**Novelty**	**Selection**
Random	Random	-	Random	-	Top k
Temporal	Temporal	-	By time	-	Top k
	SimEarliest	$tf-idf$	$cosine(t_1,t_i)$	-	Top k
Term Statistical	tfIDFSum	$tf-idf$	$\sum_{0<j<\|t_i\|} tf-idf(t_{ij})$	-	Top k
	TFIDFSum	$TF-idf$	$\sum_{0<j<\|t_i\|} TF-idf(t_{ij})$	-	Top k
Term Statistical	tfIDFSum$_N$	$tf-idf$	$\sum_{0<j<\|t_i\|} tf-idf(t_{ij})$	Similarity threshold	Top k
+Novelty	TFIDFSum$_N$	$TF-idf$	$\sum_{0<j<\|t_i\|} TF-idf(t_{ij})$	Similarity threshold	Top k
	SumBasic [10]	Language Model	$\sum_{0<j<\|t_i\|} Prob(t_{ij})$	Down-scoring terms	Top k
	Hybrid-TFIDF [12]	$TF-idf$	$Norm(t_i)\cdot\sum_{0<j<\|t_i\|} TF-idf(t_{ij})$	Similarity threshold	Top k
Cohesiveness	Centroid [11]	$tf-idf$	$cosine(centroid(T),t_i)$	-	Top k
Cohesiveness+Novelty	Centroid$_N$ [11]	$tf-idf$	$cosine(centroid(T),t_i)$	Similarity threshold	Top k

Table 1: Categorised summarisation algorithms

evaluating using a single dataset. Furthermore, there is not a generally accepted baseline for microblog summarisation, against which researchers may compare new summarisation algorithms, making it difficult to quantify the gains each new approach brings over those that came before it.

Hence, as a step towards tackling these issues, we perform a comparison of 11 microblog summarisation algorithms to determine which is the most effective. We compare the effectiveness of these algorithms for microblog summarisation using 4 Twitter datasets, and analyse their performance under both model-summary and input-summary automatic evaluation paradigms (using ROUGE [3] and SIMetrix [7], respectively). Our results confirm that summarisation algorithms that use term statistics to select tweets for inclusion into the summary are effective, supporting observations from [12], but also show that centroid-based summarisation [11] can outperform SumBasic and Hybrid TF.IDF. The remainder of this paper is organised as follows: In Section 2, we describe algorithms for microblog summarisation. We report our experimental setup in Section 3. In Section 4, we present our experimental results. Finally, Section 5 summarises our conclusions.

2 Summarisation Algorithms

Given a set of tweets, $T = \{t_1, t_2, \ldots, t_n\}$, about a topic, the task of microblog summarisation is to produce a summary composed of tweets from T, $S = \{s_1, s_2, \ldots, s_k\}$, that captures the maximum amount of essential information about the topic, within a desired summary length k (e.g. 5 or 10 tweets). Prior literature in the field of text summarisation identifies three stages that extractive summarisation algorithms typically follow [9]. First, an intermediate representation of the input documents is generated, e.g. *tf.idf* vectors. Second, each sentence is scored with respect to its preference for inclusion into the summary, where more salient or important sentences are scored highest. Third, summary sentences are selected from a ranked list (produced using the scored sentences), either by simply selecting the top k sentences for a desired summary length, or employing a redundancy filter (e.g. based on a cosine similarity threshold to previously selected sentences). We use a similar characterisation to describe the approaches to microblog summarisation examined in this paper, listed in Table 1.

Table 1 reports the 11 different summarisation approaches that we compare in our later experiments and the components that they are comprised of. t_i is a tweet to be

ranked and t_{ij} is a term in t_i. t_1 is the earliest tweet in the timeline. $cosine()$ returns the cosine similarity between two tweets. $tf-idf()$ returns the score for the term t_{ij} using the classical tf-idf weighting model. $TF-idf()$ on the other hand, returns the classical tf-idf score, with the exception that the tf component is calculated over the whole set of input tweets (with all tweets combined into a virtual document), rather than just the frequency of t_{ij} in t_i. $centroid()$ is a pseudo tweet calculated as the $tf-idf$ centroid of all tweets in the input tweet set T. $Norm()$ is a short text normalisation factor designed to avoid biasing toward longer tweets [12].

Furthermore, based upon how the different algorithms select tweets for inclusion into the summary, Table 1 also provides a categorisation of the different algorithms into six broad classes, namely: Random, Temporal, Term Statistical-Only, Term Statistical+Novelty, Cohesiveness and Cohesiveness+Novelty. We use this categorisation in our later experiments to characterise which types of algorithm are the most effective for microblog summarisation. In the next section, we describe our experimental setup, including the datasets and measures we use to evaluate microblog summarisation.

3 Experimental Setup

Evaluation Metrics: We evaluate the effectiveness of summaries, produced under each of the summarisation algorithms, using evaluation metrics from the literature: ROUGE-1 Recall; ROUGE-1 Precision; ROUGE-1 F-score; Jensen-Shannon Divergence; and Fraction of Topic Words. These metrics are implemented within the ROUGE[3] [3] and SIMetrix[4] [7] automatic summarisation evaluation tool-kits. We note, ROUGE evaluation requires a gold-standard, whereas evaluation using SIMetrix (Jensen-Shannon Divergence, and Fraction of Topic Words) does not require human authored gold-standard reference summaries (i.e. SIMetrix permits *model free* summary evaluation). We briefly describe each of the automatic summarisation evaluation metrics below:

ROUGE-N is an n-gram similarity measure between two pieces of text, from which precision, recall and f-scores are derived. In our experiments, we use ROUGE-1, which measures uni-gram overlap between a reference summary (model) and the automatically generated summary (peer) we wish to evaluate. ROUGE-1 is commonly used to measure effectiveness of microblog summarisation, due to its reported agreement with manual evaluation for short summaries [5].

Jensen-Shannon Divergence (JSD) is a measure of two probability distributions over words: the text of the original document and the text of the summary being evaluated. Low divergence [6] from the input document(s) by the produced summary is taken as a signal of an effective summary.

Fraction of Topic Words (FoTW) measures the quotient of topic words (or topic signatures [4]) of the input document(s) present in the produced summary. Effective summaries contain more topic words (from the input) in the produced summary text.

Evaluation Datasets: To compare the different microblog summarisation algorithms, discussed in the previous section, we use four microblog summarisation datasets to

[3] `http://www.berouge.com`

[4] `http://homepages.inf.ed.ac.uk/alouis/IEval2.html`

Dataset	Source	Number of Topics	Avg. Number of Tweets Per topic	Gold Standard Summaries
trending-topics-2010 (50)	Crawled via the Twitter API	50	100	✗
trending-topics-2010 (25)	Crawled via the Twitter API	25	100	✔
twitter-topics-2011/12	TREC Microblog Track 2011/12	50	167	✗
trending-topics-2014	Crawled via the Twitter API	50	100	✗

Table 2: The four tweet datasets used and their statistics.

ensure that our results are generalisable. Each dataset is comprised of sets of tweets, where each set contains tweets about Twitter trending topics or events being discussed on Twitter. Per dataset, each topic has an associated set of relevant tweets, T, which are to be summarised (i.e. the tweets are the input to the summarisation algorithms). Table 2 gives information about the four datasets, and we describe each in turn below:

trending-topics-2010 (50/25) – This dataset was obtained from Sharifi *et al.* [12]. It consists of tweets from 50 trending topics collected from the Twitter API during 2010. Notably, this dataset contains ROUGE gold-standard summaries (of length 4 tweets) for 25 of the 50 topics. As such, in our later experiments, we count this as two datasets: 'trending-topics-2010 (50)' that contains all 50 topics; and 'trending-topics-2010 (25)' that contains only the 25 topics with a gold-standard. Tweet timestamps were not provided with this dataset, hence temporal ranking approaches cannot be tested using them.

twitter-topics-2011/12 – We use a subset of the Tweets2011 corpus from the TREC Microblog track [8], taking only tweets judged relevant to the topics by NIST assessors. Ordering the collection by the number of relevant tweets per topic, we take the first 50 topics with the most tweets. The tweets are from late January to early February 2011.

trending-topics-2014 – For this dataset, we poll the Twitter API for tweets about 50 trending topics (trends in the United Kingdom). We remove non-English tweets, subsequent tweets from the same user, and filter re-tweets and near-duplicate tweets (Levenshtein distance < 5). The tweets are from late January to early February 2014.

Configuration: For both SIMetrix and ROUGE, we evaluate with stopwords removed and Porter stemming applied, to obtain a more accurate picture of textual similarity. Random performance is averaged over 10 runs. When reporting JSD and FoTW, we evaluate with a summary length of 5 tweets. When reporting ROUGE-based metrics, we evaluate at summary length 4, such that the gold-standard summaries and output summaries are the same length. Parameters within each approach are trained using a 5-fold cross validation within each dataset.

4 Results

In this section, we investigate which of the different summarisation algorithms, discussed in Section 2, are the most effective for the task of microblog summarisation. Table 3 reports the performance of each of the 11 summarisation algorithms, in terms of JSD and FoTW for all four datasets, then including ROUGE-1 Recall, Precision and F_1 for the trending-topics-2010 (25) dataset. The best performing approach under each measure/dataset pair is highlighted in bold. If two of the best approaches offer similar performances then both are highlighted. From Table 3, we observe the following.

First, comparing each approach to the random baseline, we see not all approaches outperform it. In particular, the temporal approaches (those that rank by time) and the

	SIMetrix-only						SIMetrix and ROUGE				
Approach	trending-topics-2010 (50)		twitter-topics-2011/12		trending-topics-2014		trending-topics-2010 (25)				
	JSD	FoTW	JSD	FoTW	JSD	FoTW	JSD	FoTW	Recall	Precision	F_1
Random	0.3025	0.2636	0.2653	0.2961	0.2822	0.3072	0.3147	0.2236	0.3436	0.3020	0.3149
Temporal	-	-	0.2850∗	0.2660	0.3084∗	0.2848	-	-	-	-	-
SimEarliest	-	-	0.2739	0.2556∗	0.2788	0.2944	-	-	-	-	-
tfIDFSum	0.3503∗	0.2705	0.2997∗	0.3659∗	0.3499∗	0.2625∗	0.3725∗	0.1929	0.3054	0.1797∗	0.2212∗
TFIDFSum	0.3079	0.3649∗	0.2635	0.4360∗	0.3015	0.3481∗	0.3217	0.2997∗	0.3915	0.2289∗	0.2835
tfIDFSum_N	0.3451∗	0.2784	0.3221∗	0.2554∗	0.3446∗	0.2712∗	0.3694∗	0.1727∗	0.2401∗	0.1827∗	0.1959∗
TFIDFSum_N	0.2966	0.3936∗	0.2519	0.3845∗	0.2720	0.4171∗	0.3168	0.3140∗	0.4023	0.2357∗	0.2921
SumBasic [10]	**0.2526**∗	0.3176∗	**0.2180**∗	0.3449∗	**0.2354**∗	0.3791∗	**0.2512**∗	0.2581	0.3787	**0.4596**∗	**0.4022**∗
Hybrid-TFIDF [12]	0.2892	0.3353∗	0.2472∗	0.3825∗	0.2628∗	**0.4223**∗	0.2907	0.2876∗	0.3911	0.3665∗	0.3707
Centroid [11]	0.2755∗	0.3282∗	0.2519	0.2995	0.2715	0.3057	0.2835∗	0.3066∗	0.3906	0.2912	0.3237
Centroid_N [11]	**0.2572**∗	**0.4202**∗†	**0.2143**∗	**0.4008**∗†	**0.2303**∗	**0.4325**∗†	**0.2657**∗	**0.3847**∗†	**0.4572**∗	0.3197†	0.3702

Table 3: Microblog summarisation performance using SIMetrix and ROUGE. For JSD, lower is better. For FoTW and ROUGE, higher is better. '-' denotes that the approach could not be tested on that dataset due to a lack of tweet timestamps. ∗ denotes statistical significance from random. † denotes statistically significant improvements over SumBasic by Centroid_N. Statistical significance is computed using the t-test, with $p < 0.05$.

tfIDFSum approaches produce less effective summaries than the random baseline (in some cases by a statistically significant margin, denoted ∗). For the case of the temporal approaches, this can be explained in terms of the distribution of informative information over time. By selecting tweets either by time or with respect to their similarity with the earliest tweet, informative tweets that were posted later are unlikely to be selected. The poor performance of tfIDFSum, and its novelty-enhanced version tfIDFSum_N, highlights the lack of discriminative information provided by the tf component in the microblog domain, supporting observations in [1].

Next, we compare the random baseline with the remaining approaches under the SIMetrix measures (JSD and FoTW) for the four datasets. For JSD, SumBasic and Centroid_N are the highest performing, i.e. the language model of the summaries produced by these systems diverge the least from the language model of the input tweet set T. Meanwhile, under FoTW, Centroid_N is the highest performing, i.e. the summaries produced by this system cover the largest number of important topic words. Comparing these results to the only other recent study of summarisation systems for use on microblogs [12], we observe the following. First, the high performance of term-statistic-based SumBasic approach is expected, since it was previously been shown to be one of the top three systems tested in [12]. Second, the high performance of Centroid_N which focuses on cohesiveness and novelty is surprising, since its clustering approach is similar to the classical MEAD summarisation system that was previously reported to perform poorly (it was ranked 7th out of 10 in [12]). Third, we see that the Hybrid-TFIDF approach, previously reported to be one of the best summarisation approaches is consistently outperformed by the SumBasic algorithm under JSD and ROUGE-1 Precision, and by the Centroid_N algorithm under JSD, FoTW and ROUGE-1 Recall.

Finally, comparing the best approaches, i.e. SumBasic and Centroid_N under the ROUGE metrics (Precision, Recall and F_1), we observe that these approaches perform well under different metrics. In particular, SumBasic performs well under precision, while Centroid_N performs well under recall. This indicates that SumBasic is producing more concise summaries, while Centroid_N's summaries tend to better cover the information in the gold-standard.

5 Conclusions

Effective summarisation of social media and user-generated content is an important research problem, since there are many use-cases where such sources can provide up-to-date information to end users. However, as a relatively new research topic, there has been little prior work comparing the effectiveness of summarisation algorithms specifically for the microblog domain. Hence, in this paper, we compared eleven different summarisation algorithms from the literature, over four microblog datasets, evaluating their effectiveness using five automatic summarisation evaluation metrics. Our results indicate that the SumBasic algorithm and Centroid-based summarisation with redundancy reduction were the most effective. As such, we recommend that future works report the performance of these algorithms as baselines.

Acknowledgements

All authors acknowledge the support of EC SMART project (FP7-287583). McCreadie, Macdonald and Ounis acknowledge the support of EPSRC project ReDites (EP/L010690/1).

References

[1] Amati, G., Amodeo, G., Bianchi, M., Marcone, G., Bordoni, F.U., Gaibisso, C., Gambosi, G., Celi, A., Di Nicola, C., Flammini, M.: FUB, IASI-CNR, UNIVAQ at TREC 2011 Microblog Track. In: Proc. of TREC '11 (2011)
[2] Kwak, H., Lee, C., Park, H., Moon, S.: What is Twitter, a Social Network or a News Media? In: Proc. of WWW '10 (2010)
[3] Lin, C.Y.: ROUGE: a Package for Automatic Evaluation of Summaries. In: Proc. of ACL '04 (2004)
[4] Lin, C.Y., Hovy, E.: The automated acquisition of topic signatures for text summarization. In: Proc. of ACL '00 (2000)
[5] Lin, C.Y., Hovy, E.: Automatic Evaluation of Summaries using N-gram Co-occurrence Statistics. In: Proc. of NAACL-HLT '03 (2003)
[6] Lin, J.: Divergence Measures based on the Shannon Entropy. IEEE Transactions on Information Theory 37(1) (1991)
[7] Louis, A., Nenkova, A.: Automatically Assessing Machine Summary Content without a Gold Standard. Computational Linguistics 39(2) (2013)
[8] McCreadie, R., Soboroff, I., Lin, J., Macdonald, C., Ounis, I., McCullough, D.: On Building a Reusable Twitter Corpus. In: Proc. of SIGIR '12 (2012)
[9] Nenkova, A., McKeown, K.: Automatic Summarization. Foundations and Trends in Information Retrieval 5(2-3) (2011)
[10] Nenkova, A., Vanderwende, L.: The Impact of Frequency on Summarization. MSR-TR-2005-101 (2005)
[11] Rosa, K.D., Shah, R., Lin, B., Gershman, A., Frederking, R.: Topical Clustering of Tweets (2011)
[12] Sharifi, B.P., Inouye, D.I., Kalita, J.K.: Summarization of Twitter Microblogs. The Computer Journal (2013)
[13] Spärck Jones, K.: Automatic Summarizing: Factors and Directions. Advances in Automatic Text Summarization (1999)
[14] Teevan, J., Ramage, D., Morris, M.R.: #TwitterSearch: a Comparison of Microblog Search and Web search. In: Proc. of WSDM '11 (2011)