On Choosing an Effective Automatic Evaluation Metric 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

Popular microblogging services, such as Twitter, are engaging millions of users who constantly post and share information about news and current events each day, resulting in millions of messages discussing what is happening in the world. To help users obtain an overview of microblog content relating to topics and events that they are interested in, classical summarisation techniques from the newswire domain have been successfully applied and extended for use on microblogs. However, much of the current literature on microblog summarisation assumes that the summarisation evaluation measures that have been shown to be effective on newswire, are still appropriate for evaluating microblog summarisation. Hence, in this paper, we aim to determine whether the traditional automatic newswire summarisation evaluation metrics generalise to the task of microblog summarisation. In particular, using three microblog summarisation datasets, we determine a ranking of summarisation systems under three automatic summarisation evaluation metrics from the literature. We then compare and contrast this ranking of systems produced under each metric to system rankings produced through a qualitative user evaluation, with the aim of determining which metric best simulates human summarisation preferences. Our results indicate that, for the automatic evaluation metrics we investigate, they do not always concur with each other. Further, we find that Fraction of Topic Words better agrees with what users tell us about the quality and effectiveness of microblog summaries than the ROUGE-1 measure that is most commonly reported in the literature.

Categories and Subject Descriptors

H.3.3 [**Information Storage & Retrieval**]: Information Search & Retrieval

Keywords

Social Media, Summarisation, Evaluation, Metrics

IIiX '14, August 26–29 2014 Regensburg, Germany

http://dx.doi.org/10.1145/2637002.2637017.

1. INTRODUCTION

Social media (e.g. Twitter[1]) is playing an increasingly important role in the dissemination of real-time news and information about current-events. However, due to the high volume and velocity of content posted to social media streams, particularly for very popular events such as the 2014 Super Bowl[2], there may be tens-of-thousands of posts published – far more than users can comfortably read. This means, within busy social streams, users may find it time-consuming to maintain a comprehensive and timely overview of events they are interested in or may miss relevant updates about an event they are following. To tackle this problem, text summarisation techniques [20, 32, 33] that have previously been shown to be effective for multi-document newswire summarisation, have been adapted to the task of microblog summarisation [4, 5, 24, 27, 30, 35].

Importantly, the effectiveness of these approaches were evaluated using automatic summarisation measures shown to be suitable for the newswire domain, most notably the venerable ROUGE-1 metric [11], which has become the defacto standard [20]. However, microblog summarisation has some notable disimilarities in comparison to classical newswire summarisation. For instance, posts are short and often use terminology specific to the platform (e.g. @mentions and #hashtags), while text quality (spelling and grammar) may be poorer than for newswire. Moreover, from an evaluation perspective, it is not clear whether the qualities that a summary should have are the same for newswire and microblog domains, e.g. we might expect timeliness to be more important when considering microblogs. Hence, we argue that the automatic measures used to assess summarisation quality may not generalise between domains.

On the other hand, there are a variety of automatic evaluation metrics available to researchers and practitioners for measuring summarisation effectiveness [11, 16]. In this paper, we aim to determine which of a subset of these metrics is the most suitable for evaluating microblog summmarisation. By doing so, we will facilitate more reliable automatic evaluation of different microblog summarisation algorithms. Indeed, effective summarisation evaluation is an active reasearch area. For instance, the 2012 WEAS[3] workshop on the evaluation of automatic text summarisation systems was focussed on addressing the problem of automatic evaluation within the newswire domain – the rationale for the workshop being that summarisation evaluation research

[1] twitter.com
[2] blog.twitter.com/2014/celebrating-sb48-on-twitter
[3] nist.gov/tac/2012/WEAS

Approach	Manual/Auto	Evaluation
Pyramid	Manual†	Content coverage of input documents
Responsiveness	Manual	Responsiveness to information need
ROUGE	Automatic†	N-gram overlap of gold-standard
SIMetrix	Automatic	Similarity to input documents

Table 1: **A taxonomy of four summarisation evaluation approaches. † indicates that a human authored gold-standard summary is required.**

was lagging behind research on developing new summarisation algorithms. However, there has been no corresponding work undertaken for microblog summarisation.

As such, we conduct an empirical study of automatic summarisation evaluation metrics for microblog summarisation. We investigate three different summarisation evaluation metrics from the literature, ROUGE-1, Jensen-Shannon Divergence and Fraction of Topic Words – as implemented within two publicly available automatic summarisation evaluation software tool kits, ROUGE [11] and SIMetrix [16]. These particular metrics are selected because ROUGE is the standard for reporting automatic summarisation evaluation results in the literature [20], while SIMetrix (Jensen-Shannon Divergence and Fraction of Topic Words) permits model-free summarisation evaluation (requiring no gold-standard).

In particular, we perform a system ranking comparison using three common extractive summarisation algorithms from the literature, namely: centroid-based [27]; SumBasic [23]; and Hybrid TF.IDF [30] across three different microblog datasets. First, we examine how each of the automatic evaluation metrics rank the three summarisation systems, to determine if they exhibit results which agree, i.e. to test whether they are measuring the same aspects of summary quality. Second, we perform a user-evaluation to generate pair-wise preference assessments [3] for the microblog summaries produced by each of the three approaches. We convert these preferences into a system ranking via the Ranked Pairs [34] Condorcet voting method [6], forming an aggregate preference ranking of systems, representing the ordering of summarisation systems by end-users. By comparing the system rankings produced by the automatic measures and end-users, we determine which metric best simulates end-user summarisation preference over microblogs.

The contributions of this paper are two-fold: First, we provide evidence that there are instances where automatic summarisation evaluation metrics do not concur, which means we are unable to determine if one summarisation algorithm is more effective than another. Second, we present results indicating that Fraction of Topic Words agrees with user preferences for the task of microblog summarisation, highlighting a useful alternative to the ROUGE-1 metric (which requires a gold-standard). This paper is organised as follows: Section 2 introduces background material regarding automatic text summarisation evaluation and metrics. We describe a methodology for determining the best automatic evaluation metric in Section 3. We report our experimental setup in Section 4, while Section 5 describes the user preferences study we use to establish a ground-truth ranking of summarisation systems. In Section 6, we present our results. Our conclusions are summarised in Section 7.

2. RELATED WORK

In this section, we discuss summarisation evaluation (Section 2.1) and the metrics used to measure summarisation effectiveness (Section 2.2) over newswire and microblogs. To evaluate the effectiveness of automatic text summarisation, there are manual and automatic procedures available. Manual evaluations using techniques such as the Pyramid method [21, 22], responsiveness[4] criteria, or user preference studies [28] involve humans comparing, labelling or ordering summaries. In contrast, automatic summarisation evaluation [11, 16] is a process by which a software tool is used to score summaries with respect to quality. We provide a taxonomy of summarisation evaluation approaches in Table 1.

Tool	Comparison to	Metric
ROUGE	Model Summary	ROUGE-1 Precision
ROUGE	Model Summary	ROUGE-1 Recall
ROUGE	Model Summary	ROUGE-1 F-score
SIMetrix	Source Documents	Jensen-Shannon Divergence (JSD)
SIMetrix	Source Documents	Fraction of Topic Words (FoTW)

Table 2: **The five automatic summarisation evaluation metrics we investigate in this paper.**

2.1 Evaluation of Summarisation

Performing manual summarisation evaluation can be expensive, as it requires human assessors. In contrast, automatic summarisation evaluation is intended to be a cheaper, faster and more reproducible alternative [20]. However, automatic evaluation is limited to examining textual similarity (for example, measuring n-gram overlap between a summary and a gold-standard), whereas manual evaluation procedures can account for semantics, grammar, coherence, readability [20] and other advanced means of determining the linguistic quality of a summary[5]. For automatic summarisation evaluation to be useful, it should agree with a manual summarisation evaluation involving real users.

Notably, a general constraint of summarisation evaluation relates to the summary length (number of lines or characters). Between topics, the summary length should remain constant, otherwise we may bias the evaluation towards systems which return longer summaries. A longer summary would have an unfair advantage over shorter summaries, under metrics which examine recall (or coverage) of summary content units, such as n-grams [20].

To date, previous research has examined the choice of automatic evaluation metric for evaluating multi-document newswire summarisation within the context of evaluation forums such as DUC and TAC [19, 25, 26]. However, no such research has been conducted within the microblog domain. Instead, current literature on microblog summarisation reuse evaluation techniques and metrics shown to be effective for summarisation evaluation over newswire (most often ROUGE variants [2, 15, 29]). Hence, in this paper, we examine automatic summarisation evaluation metrics for microblog summarisation, to determine the metric(s) that most closely agrees with manual evaluation.

2.2 Automatic Evaluation Metrics

Automatic summarisation evaluation metrics measure summarisation effectiveness, and are used to determine if one summarisation algorithm is more effective than another. We are aware of two publicly available automatic summarisation

[4] http://www-nlpir.nist.gov/projects/duc/duc2007/responsiveness.assessment.instructions

[5] http://www-nlpir.nist.gov/projects/duc/duc2007/quality-questions.txt

evaluation tool kits: ROUGE [11] (Recall-Oriented Understudy of Gisting Evaluation); and SIMetrix [16] (Summary Input similarity Metrics). ROUGE is the established tool for reporting automatic summarisation evaluation results in the literature, while SIMetrix is a more recent proposition. We describe each of these tool kits in more detail below.

ROUGE enables *model–peer* evaluations, examining the comparability of system produced summaries (peers) to one or more gold-standard summaries (models). We investigate ROUGE-1 (uni-gram overlap) as this is a commonly reported ROUGE variant for microblog summarisation, due to its reported agreement with manual evaluation for short summaries [13]. ROUGE evaluation requires gold-standard summaries, which are authored by humans (ideally experts in the domain) and based on a manual interpretation of the document(s) to be summarised. This is performed by NIST assessors for DUC/TAC evaluations. As noted in prior work, for ROUGE to be effective the model and peer summaries should be the same length to avoid biasing the reported F-score performance toward either precision or recall [30]. After human involvement to create exemplar summaries, ROUGE is fully automatic and repeatable.

SIMetrix enables *summary–input* evaluations, examining the reflection of the input document(s) by system produced summaries. Using the implementations within SIMetrix, we investigate Jensen-Shannon Divergence [14] (JSD) between input and summary, and the Fraction of Topic Words [12] (FoTW) of the input found in the summary – one measure of dis-similarity and one measure of similarity. Both Jensen-Shannon Divergence and Fraction of Topic Words have been reported to exhibit agreement with manual evaluation for newswire summarisation [16]. SIMetrix permits automatic summarisation evaluation without a gold-standard (model-free), in sharp contrast to ROUGE which requires gold-standard summaries. We list the metrics that we investigate in this paper in Table 2.

Although ROUGE and SIMetrix differ in how they compute summarisation effectiveness scores, both attempt to provide researchers and practitioners with reasonable automatic measures of summarisation effectiveness, which should mirror manual summarisation evaluation to some degree. In addition to ROUGE and SIMetrix (supporting Jensen-Shannon Divergence and Fraction of Topic Words) we are aware of other developments within the automatic summarisation evaluation literature [1, 7, 8, 9, 10]. However, we observe that ROUGE, often combined with manual evaluation, is still the accepted method for reporting summarisation results in the recent summarisation literature [2, 15, 29]. Hence, we focus on it and the model-free SIMetrix tool kit in this work. The problem that we address in this paper is the selection of an appropriate metric for microblog summarisation. The best metric is one that most closely agrees with manual summarisation evaluation.

3. METRICS COMPARISON METHOD

In this section, we describe the methodology used to determine if multiple automatic summarisation evaluation metrics exhibit agreement with manual summarisation evaluation results, for the task of microblog summarisation. We seek to establish what automatic summarisation evaluation metric is most suitable for automatically evaluating microblog summarisation, to ensure that automatic evaluation remains a useful proxy for manual evaluation over microblogs.

To determine how well a given summarisation measure acts as a surrogate for manual summary evaluation by end-users, we follow a three-stage methodology. We summarise each stage below and then describe each stage in more detail in the remainder of this section.

1. Use a number of summarisation algorithms (systems), to produce fixed-length microblog summaries given a set of topics. A topic is comprised of a set of related tweets, and optionally a model summary (gold-standard).

2. Score each summary using the automatic evaluation metrics to be tested. Aggregate the scores across topics for each metric to produce per-metric system rankings.

3. Perform a user preference study to obtain the ground-truth system ranking based on the preferences of end-users.

Generating Topic Summaries – The first stage of our methodology is to generate summaries for a variety of summarisation systems, representing example systems that we might be comparing in a classical summary evaluation. In this work, we use three summarisation approaches from the literature, namely: centroid-based [27]; SumBasic [23]; and Hybrid TF.IDF [30].

Automatic System Ranking – The second step is to determine a ranking of summarisation algorithms using the automatic evaluation measures to be tested. We examine ROUGE-1 (Recall, Precision and F-score), Jensen-Shannon Divergence, and Fraction of Topic Words. We use the absolute value of the measurements, provided by each of the metrics under consideration, to order the summarisation algorithms by their summarisation effectiveness scores. One complication may arise if the scores for any particular metric do not show agreement in effectiveness performance between summarisation algorithms. In this case, algorithms are then ranked in joint positions.

Manual System Ranking – For the third step in the methodology, we rank each summarisation system using qualitative assessments from end-users. This enables a comparison of rankings over both of the evaluation paradigms (automatic and manual). In order to establish such a ranking, a user evaluation is undertaken to establish the preferences of users with respect to which systems produce summaries that are more effective (higher quality). To derive the ground-truth ranking of summarisation systems we ask human assessors to compare summaries produced by the different systems, and judge which summarisation is the most effective.

Notably, there are three main methods we might consider to solicit manual summary effectiveness judgements from users. These are: graded judgements; ranked user preferences; and pair-wise user preferences [3]. With graded judgements, users are shown a single summary and asked to score that summary on a Likert scale, e.g. a five point scale from good quality to bad quality. Systems are then ranked by the scores they received. With ranked user preferences, each user is shown a series of summaries and is asked to rank them in order of preference, e.g. 1^{st}, 2^{nd}, 3^{rd}. Finally, under pair-wise user preferences, the user is shown only two summaries and asked which they prefer. These pair-wise assessments are then aggregated using a ranked pairs voting technique into a manual system ranking.

Pair	System A	System B
1	Centroid	SumBasic
2	Centroid	Hybrid
3	SumBasic	Hybrid

Table 3: **Pair-wise evaluation.**

Within a summarisation context, graded assessment of summaries is challenging, first because the grades must be defined in advance, and second, assessors may not agree on the meaning of each grade [3]. Meanwhile, ranked user preferences have a higher cognitive cost than pair-wise user preferences, since more than two summaries need to be assessed at once. Hence, for our later experiments we choose to use the pair-wise user preference assessment model, where each combination of the three summarisation systems are compared by users, as illustrated in Table 3.

However, as noted above, this pair-wise evaluation of user preferences results in judgements that do not provide a total ordering and as such we require a suitable voting method [17] to aggregate the preferences. Each pair-wise preference assessment can be seen as a vote for one summary algorithm over another. We wish to determine a ranking by where the most effective summarisation algorithm is ranked first, and least effective algorithm ranked last. Intuitively, the best ranked system should be that which was the most preferred, i.e. has the most votes between pairs. Hence, we use the Ranked Pairs [34] algorithm from the Condorcet [6] family of voting methods.

We compare the system rankings produced by both the automatic metrics and the user evaluation. The more similar the system rankings for a metric across topics is to the human ranking, the more effectively that measure acts as a surrogate for end-users. In the next section, we describe the experimental setup, while in Section 5, we describe our crowdsourced user preferences study. Our results are reported in Section 6.

4. EXPERIMENTAL SETUP

In this section, we describe the experimental setup used to address the following two research questions.

1. Do the automatic summarisation evaluation metrics agree which of the summarisation algorithms tested are more effective?

2. Of the automatic summarisation evaluation metrics tested, which most closely reflects the preferences collected during our user evaluation?

The remainder of this section is structured as follows. In Section 4.1, we describe the microblog summarisation algorithms (systems) used to produce summaries. Section 4.2, describes the microblog summarisation datasets. In Section 4.3, we detail the automatic summarisation evaluation metrics tested. Section 4.4 reports the measure configurations and the training regime we use.

4.1 Summarisation Algorithms

In this section, we describe the algorithms for microblog summarisation used in our later experiments. In particular, we implement three competitive extractive summarisation algorithms from the literature [23, 27, 30]. Each algorithm takes in a set of tweets and produces an automatic text summarisation consisting of a subset of the input tweets. Formally, given a set of input tweets, $T = \{t_1, t_2, \ldots, t_n\}$, the microblog summarisation task is to produce a summary, $S \in T$, composed of tweets from T, that captures the maximum amount of essential information from the set of input tweets, within a desired summary length k (e.g. five tweets). We describe each of the algorithms below.

Centroid – Centroid-based summarisation of a set of tweets is based on comparing each tweet to a centroid pseudo-tweet [27]. The centroid of the set of input tweets, $T = \{t_1, t_2, \ldots, t_n\}$, is calculated as: $centroid = \frac{LS}{N}$, where LS is the linear sum of the *tf.idf* vectors of the tweets in T, and N is the number of tweets in T. The set of input tweets are scored by cosine similarity to this centroid vector, with the highest scoring tweets are selected for summary.

$$\overset{score(t_i)}{Centroid} = Sim(centroid)$$

To reduce redundancy and promote novelty, a cosine similarity threshold is applied, selecting summary tweets only if they are sufficiently dissimilar to previously selected tweets.

SumBasic – SumBasic relies on input word frequency to determine sentence selection, as it was found that words occurring frequently in a given document are more likely to be included in a human summary [23]. SumBasic is an iterative greedy algorithm, scoring tweets based on the average probability of the words appearing in a tweet. During each iteration, the tweet containing the words with the highest probability are selected. The probability of a word, $p(w_i)$, is equal to $\frac{F}{W}$, where F is the word's frequency over all input tweets, and W is the total number of words in the set of input tweets. The average probability of words, w_i, in a tweet, t_i, is scored:

$$\overset{score(t_i)}{SumBasic} = \sum w_i \in t_i \frac{p(w_i)}{|\{w_i \mid w_i \in t_i\}|}$$

Once a tweet has been selected, for each word, w_i in the selected tweet the word's probability value is reduced: $p_{new}(w_i) = p_{old}(w_i) \times p_{old}(w_i)$. The p_{old} is the probability of the word being included in the summary, while p_{new} is the probability of the word being included in the summary for a second time. This update of word probabilities results in the next iteration being informed by the previous iteration – i.e. the probability of a word being included in a summary depends on whether the word has previously been included.

Hybrid – Another class of summarisation algorithm uses the *tf.idf* scores for terms in each tweet for ranking. Under these approaches, the score for a tweet is calculated as the sum of the *tf.idf* scores for each term contained. Tweets are represented as n-dimensional vectors, $t_i = t_{i1}, t_{i2}, \ldots, t_{in}$, of *tf.idf* scores, where each tweet is scored by summing the values of the term vector components:

$$\overset{score(t_i)}{tf-idf} = \sum_{i=1}^{n} t_i = t_{i1} + t_{i2} + \ldots + t_{in}$$

The *top-k* tweets are selected with k being the desired summary length. In a similar manner to Centroid summarisation, to limit redundancy, a cosine similarity threshold is applied such that a tweet is selected only if it is sufficiently dissimilar to previously selected tweets.

The *hybrid tf.idf* algorithm extends classical tf.idf approaches for use in the microblog domain [30]. Under this approach, the *tf* component is calculated over the whole set of input

Dataset	Source	Topics	Tweets	ROUGE	SIMetrix
microblog-track	Tweets2011	50	135	X	✓
trending-topics-2014	Twitter API	50	100	X	✓
trending-topics-2010 (50)	Twitter API	50	100	X	✓
trending-topics-2010 (25)	Twitter API	25	100	✓	✓

Table 4: **The statistics of each microblog dataset.**

tweets (with all tweets combined into a virtual document), while the *idf* component is calculated as per classical *tf.idf* (with each tweet taken as an individual document in a collection). Further, *tf.idf* length normalisation [31] is employed to avoid a bias towards longer tweets. Hybrid *tf.idf* scores a tweet, $t_i = t_{i1}, t_{i2}, \ldots, t_{in}$, as:

$$\underset{Hybrid}{score(t_i)} = \frac{t_{i1} + t_{i2} + \ldots + t_{in}}{\max[MinimumThreshold, WordsInTweet]}$$

The Centroid, SumBasic and Hybrid algorithms are representative of the state-of-the-art for extractive summarisation of microblogs [30]. We evaluate these algorithms over microblog datasets with varying characteristics, which we describe in the next section.

4.2 Microblog Summarisation Datasets

For evaluating microblog summarisation we use four microblog datasets. For each dataset, Table 4 lists the source of the tweets, the number of topics, the average number of tweets per topic, and the appropriate evaluation tools. Each microblog summarisation dataset comprises a number of topics, which are collections of tweets that are related by a common theme. For the datasets in Table 4, topics are either derived from Text REtrieval Conference (TREC) adhoc retrieval topics, or trending topics as defined by Twitter. Notably, SIMetrix *summary-input* evaluation is conducted without a gold-standard, while, ROUGE *model-summary* evaluation requires a gold-standard. Not all of the summarisation datasets we use have a gold-standard, hence in our later experiments we report ROUGE evaluation only on those datasets (see Table 4). We describe each dataset in more detail below.

microblog-track – We use a subset of the Tweets2011 corpus from the TREC Microblog track [18], taking only tweets judged relevant to the topics by NIST assessors. To make the summarisation task non-trivial, we require topics that contain as many tweets about them as possible. For this dataset, we selected the top 50 topics with the most tweets, using them for microblog summarisation. In this way, we ensure that each topic contains at least 60 tweets to be summarised (range 60–524 tweets, average 135 per topic). The tweets are from late January to early February 2011. This dataset does not have gold-standard summaries and hence is used to report Jensen-Shannon Divergence and Fraction of Topic Words only.

trending-topics-2014 – For this dataset, we poll the Twitter API for tweets about 50 trending topics in the United Kingdom. We remove non-English tweets and subsequent tweets from the same user. Additionally, we filter out retweets and near-duplicate tweets (Levenshtein distance < 5). The tweets are from late January to early February 2014. This dataset does not have gold-standard summaries.

trending-topics-2010 (50/25) – This dataset was obtained from Sharifi *et al.* [30]. It consists of tweets from 50 trending topics collected from the Twitter API during 2010. This dataset contains gold-standard summaries, of length 4 tweets, for 25 of the 50 topics. As such, in our later experiments we report Jensen-Shannon Divergence and Fraction of Topic Words, at a summary length of 5 tweets over all 50 topics, denoted **trending-topics-2010 (50)**. We then report all summarisation measures (ROUGE and SIMetrix-based) over the 25 topics that have gold-standard summaries, denoted **trending-topics-2010 (25)**.

4.3 Evaluation Metrics

We evaluate how five automatic summary evaluation metrics from the literature compare to human summary preferences. These metrics are: ROUGE-1 Recall, ROUGE-1 Precision, ROUGE-1 F-score, Jensen-Shannon Divergence and Fraction of Topic Words. We detail each below:

ROUGE-N is an n-gram similarity measure between two pieces of text, from which ROUGE 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 we wish to evaluate. ROUGE-N is defined as:

$$\text{ROUGE-N} = \frac{\sum_{S \in \{RS\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{RS\}} \sum_{gram_n \in S} Count(gram_n)}$$

where RS is one or more reference summaries (comprising the gold-standard), n is the n-gram length, 1 for ROUGE-1, $Count(gram_n)$ is the n-grams in the gold-standard, and $Count_{match}(gram_n)$ is the n-grams matching between the gold-standard and evaluated summary.

Jensen-Shannon Divergence (JSD) is a measure of two probability distributions of words: the text of the original document, P, and the evaluated summary text, Q. Low divergence from the input document(s) by the produced summary is taken as a signal of a good summary. Given two probability distributions over words, P and Q, Jensen-Shannon Divergence is defined:

$$\text{JSD}(P||Q) = \frac{1}{2}[D(P||A) + D(Q||A)]$$

where A is defined as $\frac{P+Q}{2}$ (mean distribution of P and Q).

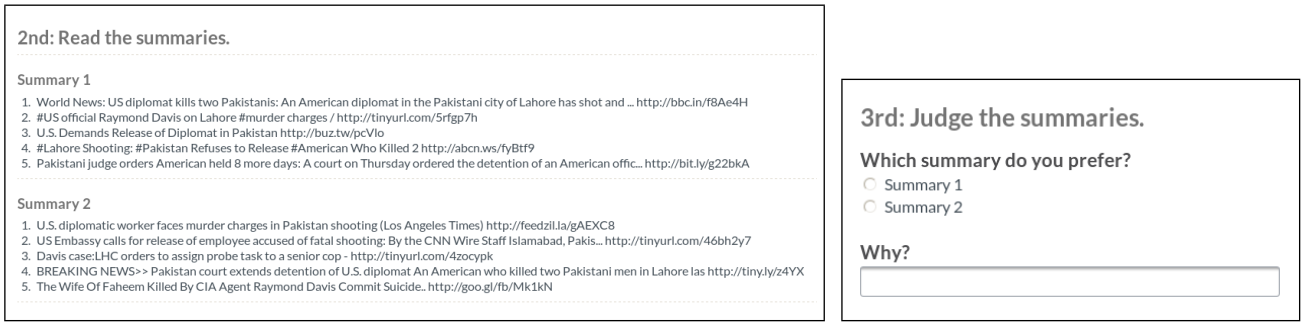
Fraction of Topic Words (FoTW) measures the quotient of topic words (or topic signatures) derived from the input documents that are also found in the summary. Topic signatures were originally proposed as a feature for summarisation [12]. Used as a summary evaluation method, the more topic words (from the input) found in the summary text, the better the summary is considered to be. Compared to a background corpus, a topic word is a textual unit (word, stemmed term, bi-gram or tri-gram) with a statistically significantly greater probability in a specific document, than within the background corpus. For all the words in a document, the ratio between hypotheses (1) and (2) is calculated: (1) This word is not a topic word, it has the same probability in the document and background corpus. (2) This word is a topic word, it has a greater probability in the document, over the background corpus. Given the ratio, λ, $-2\log\lambda$ has a χ^2 distribution. This provides a cut-off parameter, with words above this cut-off declared as topic words.

4.4 Settings & Parameters

For both ROUGE[6] and SIMetrix[7], we evaluate with stopwords removed and Porter stemming applied. The Centroid

[6] `www.berouge.com`

[7] `homepages.inf.ed.ac.uk/alouis/IEval2.html`



(a) Two microblog summaries (b) Assessment form

Figure 1: **An example of the summaries shown to users, and the assessment form used to collect judgements.**

and Hybrid summarisation algorithms have the following parameters that are trained. For the Centroid algorithm, there is a cosine similarity threshold (0-1, increments of 0.05). For the Hybrid algorithm, there is a cosine similarity threshold (0-1, increments of 0.05) and a normalisation factor (0-20, increments of 2). We use a 5-fold cross validation within each dataset to train these parameters. The loss function is Jensen-Shannon Divergence.[8] The SumBasic algorithm has no parameters to optimise. When reporting performance on the datasets without a gold-standard, the target summary length is 5 tweets. On the dataset with a gold-standard (trending-topics-2010 (25)), the target summary length is 4 tweets. This is a limitation of ROUGE, i.e. the summary length must equal the length of the gold-standard.

5. USER PREFERENCES STUDY

In this section, we describe the operationalisation of the metrics comparison method in Section 3. In order to determine a manual ground-truth ranking of summarisation algorithms, we ask users to judge the quality of the summaries produced by each algorithm. For the summarisation algorithms, Centroid, SumBasic and Hybrid, a per topic, pairwise evaluation is conducted over the microblog datasets. The preferences for one summarisation algorithm over another, expressed by users, are used to produce a ranking of the summarisation algorithms.

For each topic in a dataset, we have three pairs to evaluate, as shown earlier in Table 3. The desired outcome of the pairwise evaluation is a ranking, $x < y < z$ (i.e. 1^{st}, 2^{nd}, 3^{rd}), of the summarisation algorithms, as determined by user's preferences. Given three pairs, we have six possible rankings of the summarisation algorithms, as shown below:

Ordering	**Possible outcomes** (rankings of summarisation algorithms)
h < sb < c	Centroid is better than SumBasic, SumBasic is better than Hybrid.
sb < h < c	Centroid is better than Hybrid, Hybrid is better than SumBasic.
h < c < sb	SumBasic is better than Centroid, Centroid is better than Hybrid.
c < h < sb	SumBasic is better than Hybrid, Hybrid is better than Centroid.
sb < c < h	Hybrid is better than Centroid, Centroid is better than SumBasic.
c < sb < h	Hybrid is better than SumBasic, SumBasic is better than Centroid.

Whichever is the most popular ordering indicated by manual evaluation, is then compared to a ranking of the summarisation algorithms established via automatic evaluation. In this paper, we use the CrowdFlower[9] crowdsourcing platform to recruit workers and gather manual judgements.

[8]To reduce the variables within the experiment we do not examine other potential loss functions.

[9]`crowdflower.com`

The intended outcome of the experiment is to establish whether automatic evaluation metrics agree with manual summarisation effectiveness judgements. We seek to determine if automatic evaluation metrics can accurately differentiate between good and bad summaries, as determined by real users (the main desirable property of an automatic summarisation evaluation metric). In Section 5.1, we describe the crowdsourced experiment, and in Section 5.2, we describe the quality control methods implemented.

5.1 Crowdsourcing Job Description

Per dataset, per topic, we display pairs of summaries and ask workers to indicate which summary they think is better. While completing the assignment, workers are asked to read two summaries. Workers may optionally read the original set of tweets that the summaries were derived from, but this is not a requirement of the task. The question asked of the workers is "Which summary do you prefer?". We do not guide workers by providing any specific criteria by which they must judge the summaries, leaving the decision open-ended. A second question asked, "Why?", solicits a free-text reason for the workers's preference in the first question.

Workers are paid $0.05 per unit of work. A unit of work consists of three judgements. A judgement involves reading two summaries, indicating which summary is best, and providing a free-text reason. The forms workers interact with to complete assessments are shown in Figure 1 (a) and (b).

5.2 Crowdsourcing Quality Control

To obtain high-quality judgements from the crowd, performed quality assurance as follows. Test questions were configured for the job, requiring workers to correctly assess that a sample Centroid summary is better than a summary consisting of random *lorem ipsum* text. Workers were trained on these test questions prior to task entry (quiz mode), and test questions are inserted throughout the task by the CrowdFlower platform. Workers who fail to answer test questions correctly have their judgements eliminated from the pool. Additionally, workers must pass a simple general knowledge quiz to submit their work for a task, where they are asked to identify the presidents of countries (UK, US, Russia). These measures attempt to stop workers who are not paying attention to the job, from erroneously influencing the results.

We restrict worker location to UK, US, Canada, Austrailia and New Zealand (English speaking nations), as the tweets from each of our microblog datasets are in English, and we do not wish judgements from workers who are not able to

Experiment	Agreement	Assessed
microblog-track	77.78	43/50 (86%)
trending-topics-2014	78.05	50/50 (100%)
trending-topics-2010 (50)	76.77	43/50 (86%)
trending-topics-2010 (25)	77.35	22/25 (88%)

Table 5: **Crowdsourced user preferences statistics.**

fully comprehend the microblog summaries they are asked to read. As we seek quick completion of a relatively simple task, we permit CrowdFlower level 1 contributors (all workers) to complete our task for efficiency reasons. However, a maximum of twelve judgements per worker is accepted, to reduce the effects of more active crowdworkers skewing results towards their repeatedly similar judgements. Further, the ordering of summary pairs is rotated to mitigate selection bias, and each summary pair per topic is judged three times by different workers. The final summary preference pair for each topic is resolved via a majority vote.

Table 5 shows statistics from the crowdsourced manual evaluation. For the pair-wise evaluation over each dataset, we show the CrowdFlower provided assessor agreement, the number of topics for which we obtained complete judgements (all pairs judged three times), and the number of topics that we could derive total $x < y < z$ orderings (no tie-breaks), from pair-wise preferences.

Under the "agreement" column, we show the breakdown in figures giving an average agreement of 77.5% over the experiments. With pair-wise evaluation, random agreement would be 50%, so we obtain reasonable inter-assessor agreement using the CrowdFlower platform. Under the "assessed" column, we see a reduction from the number of topics in the dataset due to noise in the crowdsourcing platform. Although the CrowdFlower platform forces all workers to submit complete assessments, the platform removes judgements from the pool of those workers who fail the test questions. We remove topics containing such partial judgements.

6. RESULTS & ANALYSIS

In this section, we present our results from the automatic and manual summarisation evaluations. We investigate how the automatic summarisation metrics relate to each other in Section 6.1 and then compare how each metric agrees with our user evaluation in Section 6.2, We then analyse and discuss the evaluation measures in more detail in Section 6.3.

6.1 Do the automatic evaluation metrics agree?

We begin by evaluating whether the automatic summarisation evaluation metrics, namely: Jensen-Shannon Divergence, Fraction of Topic Words and ROUGE-1, exhibit approximately similar results that do not conflict with each other, i.e. do they agree and tell us the same thing? For a range of summarisation algorithms, over a number of tweet datasets, we would expect that the automatic evaluation results for any given metric are reflected by the other metrics. For example, if the SumBasic algorithm performs well under Jensen-Shannon Divergence, we would expect SumBasic to also perform well under the Fraction of Topic Words metric – if the metrics are measuring the same aspects of summarisation effectiveness. However, if Jensen-Shannon Divergence does not report the same result as Fraction of Topic Words, then these metrics are measuring different aspects of summarisation effectiveness.

Table 6 reports the mean Jensen-Shannon Divergence (JSD) and Fraction of Topic Words (FoTW), over three microblog datasets, for three summarisation algorithms. Over each of the microblog-track, trending-topics-2014, and trending-topics-2010 (50) microblog datasets, we may only derive JSD and FoTW results as there are no gold-standard summaries available (required by ROUGE). For JSD results, lower is better (less divergence). For FoTW results, higher is better (more topic words). Table 6 also reports the ranking of summarisation algorithms produced by the automatic evaluation metrics. In the case where metrics show no statistically significant difference (t-test, $p \leq 0.05$) between the systems, we rank the algorithms in joint position (for example Centroid and SumBasic score very similarly under JSD over all three datasets). From Table 6, we observe that the results for JSD and FoTW are consistent over the three datasets, for example: FoTW always ranks the summarisation algorithms as: Centroid, Hybrid then SumBasic. Meanwhile, JSD always ranks the algorithms differently, i.e. Centroid/SumBasic (joint) and then Hybrid. Furthermore, we see that Centroid generally scores best, under both JSD and FoTW, being ranked 1^{st} or 2^{nd} over the three datasets. However, we observe discrepancies within the results over the three datasets. FoTW consistently places SumBasic in 3^{rd} place, while JSD ranks SumBasic as 1^{st} or 2^{nd}. For Hybrid, JSD places this algorithm in 3^{rd} place, while FoTW consistently ranks Hybrid 2^{nd}. Under JSD, Centroid and SumBasic compete for 1^{st} place, but FoTW splits the two into 1^{st} and 3^{rd} ranks. In summary, Jensen-Shannon Divergence and Fraction of Topic Words do not to produce the same ranking of summarisation systems.

Table 6 also reports Jensen-Shannon Divergence (JSD), Fraction of Topic Words (FoTW), and ROUGE-1 Recall, Precision and F-score, over the trending-topics-2010 (25) dataset. For this dataset we can calculate ROUGE-1 Recall, Precision and F-score metrics, in addition to JSD and FoTW, as there are gold-standard summaries available. From Table 6, we observe that each automatic evaluation metric produces a different system ranking. For instance, JSD does not agree with FoTW, nor any ROUGE-1 metric. Further, ROUGE-1 F-score places SumBasic 1^{st}, while FoTW places SumBasic 3^{rd}. Overall, from results in Table 6, we conclude the metrics exhibit marked disagreement on the ranking of systems.

To examine this behaviour in more detail, Table 7 reports the Spearman's correlation between the scores produced by each metric across the topics in all datasets[10]. High correlation values indicate that the metrics are functionally equivalent, while low correlations indicate that each is measuring a different aspect of summary quality. From Table 7, we observe that the two summary-input metrics (JSD and FoTW) have a moderate negative correlation, and observe that the ROUGE-1 metrics (Recall, Precision, F-score) are highly correlated with each other. Comparing the summary-input metrics (JSD and FoTW) to the model-peer metrics (ROUGE-1 Recall, Precision and F-score), we observe negligible correlation between JSD and ROUGE-1. However, we do observe moderate correlation between FoTW and ROUGE-1, with FoTW correlating most with ROUGE-1 Recall, over ROUGE-1 Precision and F-score.

[10] JSD and FoTW are calculated over all four datasets (n=474), while ROUGE-1 Recall, Precision and F-score are calculated only over the trending-topics-2010 (25) dataset (n=66), i.e. the only one with a gold-standard.

	microblog-track				trending-topics-2014				trending-topics-2010 (50)			
	JSD	**Rank**	**FoTW**	**Rank**	**JSD**	**Rank**	**FoTW**	**Rank**	**JSD**	**Rank**	**FoTW**	**Rank**
Centroid	**0.2143**	$1^{st}/2^{nd}$	**0.4008**	1^{st}	**0.2303**	$1^{st}/2^{nd}$	**0.4325**	1^{st}	0.2572	$1^{st}/2^{nd}$	**0.4202**	1^{st}
SumBasic	0.2180	$1^{st}/2^{nd}$	0.3449	3^{rd}	0.2354	$1^{st}/2^{nd}$	0.3791	3^{rd}	**0.2526**	$1^{st}/2^{nd}$	0.3176	3^{rd}
Hybrid	0.2472	3^{rd}	0.3825	2^{nd}	0.2628	3^{rd}	0.4223	2^{nd}	0.2892	3^{rd}	0.3353	2^{nd}

	trending-topics-2010(25)									
	JSD	**Rank**	**FoTW**	**Rank**	**Recall**	**Rank**	**Precision**	**Rank**	**F-score**	**Rank**
Centroid	0.2657	$1^{st}/2^{nd}$	**0.3847**	1^{st}	**0.4572**	1^{st}	0.3197	3^{rd}	0.3702	$2^{nd}/3^{rd}$
SumBasic	**0.2512**	$1^{st}/2^{nd}$	0.2581	3^{rd}	0.3787	$2^{nd}/3^{rd}$	**0.4596**	1^{st}	**0.4022**	1^{st}
Hybrid	0.2907	3^{rd}	0.2876	2^{nd}	0.3911	$2^{nd}/3^{rd}$	0.3665	2^{nd}	0.3707	$2^{nd}/3^{rd}$

Table 6: **Automatic summarisation results over the four datasets, additionally showing system rankings. Two system share the same rank if the difference in their performance is not statistically significant (t-test $p \leq 0.05$).**

	JSD	**FoTW**	**Recall**	**Precision**	**F-score**
JSD	**1.0**	-0.4210*	-0.0372	-0.0211	-0.0508
FoTW	-0.4210*	**1.0**	0.6201*	0.4354*	0.5383*
Recall	-0.0372	0.6201*	**1.0**	0.8878*	0.9556*
Precision	-0.0211	0.4354*	0.8878*	**1.0**	0.9762*
F-score	-0.0508	0.5383*	0.9556*	0.9762*	**1.0**

Table 7: **Spearman's correlation between summarisation evaluation metrics, across all datasets. * denotes a significant correlation (p≤0.05)**

This matches the similarity noted between FoTW and ROUGE-1 Recall from results in Table 6. From Table 7, we conclude that the measures JSD and ROUGE-1 are not correlated, and FoTW and ROUGE-1 are only moderately correlated.

To answer our first research question, we observed discrepancies between results for the Centroid, SumBasic and Hybrid algorithms, across the microblog datasets. Of the metrics investigated, used to report summarisation effectiveness, only FoTW and ROUGE-1 appear to be exhibiting similar results. JSD does not exhibit comparable results to FoTW or the ROUGE-1 metrics, while FoTW does not exhibit comparable results to ROUGE-1. This result demonstrates that these evaluation metrics do not agree on which system is the more effective – i.e. they are measuring different aspects of summarisation effectiveness. Overall, the disagreement observed between the metrics means that without further information about which metric is the most effective, it is not possible to confidently rank the summarisation systems using them. Hence, in the next section, we investigate which automatic summarisation evaluation metric exhibits the most agreement with our manual summarisation evaluation, i.e. such that we can determine which metric is the most effective.

6.2 Which metric matches manual evaluation?

To answer our second research question, we investigate which automatic evaluation metric(s), ROUGE-1, Jensen-Shannon Divergence, and Fraction of Topic Words, best agrees with observations from manual evaluation, i.e. which metric better reflects summarisation effectiveness judgements from real users? For example, if ROUGE-1 reports results that agree with manual evaluation, while Jensen-Shannon Divergence and Fraction of Topic Words do not, then the ROUGE-1 metric provides a more accurate estimate of what users think about summarisation effectiveness. We would expect that automatic evaluation metrics exhibit at least some agreement with manual evaluation, but we do not know which particular metric most closely correlates with manual evaluation for the task of microblog summarisation.

Table 8 reports the number of topics for which users preferred each ordering of systems under the Ranked Pairs Condorcet voting method, in addition to the number of times that each system was ranked first. From Table 8, we observe that Centroid is consistently preferred by the manual assessors. For example, for the top two most frequent outcomes over all the datasets, users preferred the Centroid-based summaries. Further, Table 8 also shows that the system ranking provided by manual assessment directly matches the system ranking established by automatic evaluation under FoTW (see Table 6). We also observe that the Centroid and SumBasic algorithms are split by user votes by a wide margin e.g. 20 votes for Centroid and 5 votes for SumBasic on the microblog-track dataset. Over the microblog-track and trending-topics-2014 datasets, Hybrid is ranked 2^{nd}, between Centroid and SumBasic. However, Hybrid is much closer to Centroid for the trending-topics-2010 (25) dataset and almost equal to SumBasic over trending-topics-2010 (25), indicating that the performance of these systems is closer. From the results in Table 8, we conclude that Fraction of Topic Words is more closely correlated with manual evaluation than Jensen-Shannon Divergence, ROUGE-1 Recall, ROUGE-1 Precision, or ROUGE-1 F-score. In answer to our second research question, of the automatic evaluation metrics investigated, FoTW exhibits the most agreement with manual summarisation evaluation.

6.3 Discussion

Given that we have established that Fraction of Topic Words (FoTW) most closely agrees with manual evaluation, for the task of microblog summarisation, we now analyse and discuss why this is the case. We begin by examining some example summaries, to see why Fraction of Topic Words performs well. Below, we show the summaries produced for the Centroid and SumBasic systems for the topic "high taxes" within the microblog-track dataset. The terms that Fraction of Topic Words uses to score each summary are highlighted.

Centroid

(1) Wait, wait...did Obama just say we need to cut corporate tax rates & we have one of the highest corporate tax rates in the world?! #sotu

(2) ObamaCare Flatlines: ObamaCare Taxes Home Sales - Clobbers Middle-Class Americans - Blog - GOP.gov http://bit.ly/9fyRgM

(3) 750,000 'to pay higher tax rate': Three-quarters of a million more people are set to become higher rate taxpayers in April, says the ...

(4) Generrational warfare on the news... i find it increasingly hard to see why an increased higher rate of income tax is deemed unfair?

(5) The GOP Party of the Filthy Rich; its an outrage to let the rich off the hook in paying taxes. http://bit.ly/dG1U7q

SumBasic

(1) Wait, wait...did Obama just say we need to cut corporate tax rates & we have one of the highest corporate tax rates in the world?! #sotu

(2) Flood levy in addition to the carbon tax & mining tax. What next - a breathing tax?

(3) Unseen taxes you pay everyday... http://yhoo.it/gJDQjX

(4) Thousands More To Pay Higher-Rate Income Tax: Three-quarters of a million people will have to... http://goo.gl/fb/3Xf7t

(5) http://ping.fm/p/7ahpp - Higher tax rate to hit 750,000 more people, says IFS

microblog-track		trending-topics-2014		trending-topics-2010 (50)		trending-topics-2010 (25)	
Ordering	Freq.	Ordering	Freq.	Ordering	Freq.	Ordering	Freq.
sb < h < c	13	sb < h < c	12	sb < h < c	11	h < sb < c	08
sb < c < h	09	sb < c < h	11	sb < c < h	08	sb < h < c	07
h < sb < c	07	h < sb < c	08	c < h < sb	06	sb < c < h	02
h < c < sb	04	c < sb < h	03	c < sb < h	06	c < h < sb	01
c < sb < h	01	h < c < sb	02	h < sb < c	05	h < c < sb	01
c < h < sb	00	c < h < sb	01	h < c < sb	00	c < sb < h	01

	microblog-track		trending-topics-2014		trending-topics-2010 (50)		trending-topics-2010 (25)	
	1st Places	Rank	1st Places	Rank	1st Places	Rank	1st Places	Rank
Centroid	20	1st	20	1st	16	1st	15	1st
SumBasic	04	3rd	03	3rd	06	3rd	02	3rd
Hybrid	10	2nd	14	2nd	14	2nd	03	2nd

Table 8: **User preferences over each of the four datasets. The upper sub-table reports the number of times each total ordering (ranking) of systems was observed, while the lower sub-table reports the number of times each system was ranked first under the Ranked Pairs Condorcet voting method. For each dataset, we show the frequency of each $x < y < z$ ordering.**

For this topic, our users preferred the Centroid summary over the SumBasic summary. Meanwhile, both Jensen-Shannon Divergence and Fraction of Topic Words agree. The reasons users gave for preferring the Centroid summary to the SumBasic summary in our user evaluation (see Section 5), for this topic, were that it contained a better level of detail and a higher number of links. From the above summaries, we observe that the Centroid summary contains a higher proportion of (highlighted) topic words, indicating that it is capturing the additional detail in that summary. Indeed, the highlighted topic words in the above microblog summaries are give a brief overview of what is being discussed in the tweets. For instance, we can easily see that this topic is about the Obama administration's tax policy. From simply reading the highlighted topic words alone, we get a sense of what the topic is about, which indicates that topic words are very informative.

Furthermore, from the reasons given we know that the users are considering the links within the tweets when making their assessments, which are not captured by the automatic measures. Hence, a direction for future work would be to extend the automatic metrics to incorporate evidence from the hyperlinked documents.

Next, we examine an example topic where the automatic summarisation measures do not agree with manual assessments. In particular, below we show the SumBasic and Hybrid summaries for the topic "A-Rod", which is about a baseball player's sporting record of home runs.

SumBasic
(1) A-Rod homers in third straight game http://bit.ly/168LMB
(2) a-rod and arod are trending.
(3) @StaceGots Jeter and A-Rod.
(4) Go Yankees!!! A-Rod is unstoppable!!!

Hybrid
(1) #A-Rod is kobe
(2) watching a-rod tie howard and gehrig's postseason rbi streak record. howard also tied gehrig's 70+ year old record just this year.
(3) AROD and A-ROD both trending now... Interesting
(4) A-Rod homers in third straight game http://bit.ly/168LMB #MLB #Baseball

The users preferred the Hybrid summary over the SumBasic summary for this topic. Fraction of Topic Words and ROUGE-1 Recall both agree with the user assessment. However, Jensen-Shannon Divergence, ROUGE-1 Precision and ROUGE-1 F-score disagree (SumBasic is ranked above Hybrid under these metrics). For this topic, reasons provided by manual assessors for preferring the Hybrid summary over the SumBasic summary again relate to informativeness.

As we can see from the example summaries, Fraction of Topic Words correctly ranks the Hybrid summary over the SumBasic summary, as it contains a higher proportion of topic words. On the other hand, Jensen-Shannon Divergence fails to correctly rank the summaries. From further analysis, this is because the raw text of the Hybrid summary diverges markedly (0.3109 JSD) from the term distribution of the input tweets, i.e. it contains rare information (that Alex Rodriguez tied with two other baseball players RBI streak records). This compares to a lower JSD value of 0.2447 for the SumBasic summary. This result highlights a case where Jensen-Shannon Divergence can fail, i.e. when key information that we want to capture is contained in only a few of the input tweets. Considering the ROUGE-1 metrics, ROUGE-1 Recall correctly ranks the summarisation systems, agreeing with manual evaluation. However, ROUGE-1 Precision does not agree with the manual summarisation evaluation. The ROUGE-1 Precision score for the SumBasic summary is 0.6177, compared to Hybrid's precision score of 0.3438. Importantly, the Hybrid summary contains a long and informative tweet (number 2 above), which contains information that does not feature in the gold-standard for this topic[11]. This means that the ROUGE metrics cannot detect that the SumBasic summary is missing the information contained within this clearly topical and relevant tweet. This result shows that if ROUGE variants are to be reported for microblog summarisation, then it is critical that the human-authored gold-standard contain all of the key information.

Finally, to identify the key summary aspects that an effective microblog summarisation metric should capture, we report the proportion of reasons given for selecting one summary over another, from pair-wise evaluation. In particular, we manually categorised the reasons provided by crowdsourced workers into 5 categories, as shown in Table 9. From Table 9, we observe that users tell us the two main criteria a good microblog summarisation metric should cover are informativeness and readability. However, summary length, presence of sentiment and tweet-specific features such as hyperlinks contained can also be important aspects to consider.

[11]The gold-standard is omitted due to space constraints.

Category	Weight	Examples
informativeness	43.5%	*"more descriptive","better details".*
readability	35.1%	*"easier to understand","better grammar".*
length	9.1%	*"more concise","to the point".*
sentiment	7.3%	*"more excitement","sounds fun".*
tweet-specific	5.0%	*"more hashtags","better news links".*

Table 9: **Samples of crowdworker's stated reasons for summary preference, organised into five categories.**

7. CONCLUSIONS

In this paper, we examined automatic summarisation evaluation metrics for the task of microblog summarisation. We showed that there are instances where the ROUGE-1, Jensen-Shannon Divergence, and Fraction of Topic Words automatic evaluation metrics do not agree with each other, and it would appear they are measuring different aspects of microblog summarisation effectiveness. Moreover, a user preferences study which examines the agreement between automatic and manual summarisation evaluation, indicates that (of the metrics tested) Fraction of Topic Words agrees most often with manual summarisation evaluation. Furthermore, we have shown that the automatic evaluation metric Fraction of Topic Words appears to perform most effectively (agrees with manual summarisation evaluation) because the underlying topic words being used often effectively summarise the key information contained within the microblog topics. Finally, also from the user preferences study, we have identified that informativeness and readability are the two key aspects that a good microblog evaluation metric needs to capture. In conclusion, we found that Fraction of Topic Words better agrees with what users tell us about the quality and effectiveness of microblog summaries than Jenson-Shannon Divergence and the ROUGE-1 metrics tested.

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] E. Amigó, J. Gonzalo, and F. Verdejo. The Heterogeneity Principle in Evaluation Measures for Automatic Summarization. In *Proc. of WEAS*, 2012.

[2] E. Baralis, L. Cagliero, S. Jabeen, A. Fiori, and S. Shah. Multi-document Summarization based on the Yago Ontology. *Expert Systems with Applications*, 40(17), 2013.

[3] B. Carterette, P. N. Bennett, D. M. Chickering, and S. T. Dumais. Here or There: Preference Judgments for Relevance. In *Proc. of ECIR*, 2008.

[4] D. Chakrabarti and K. Punera. Event Summarization using Tweets. In *Proc. of ICWSM*, 2011.

[5] F. Chua and S. Asur. Automatic Summarization of Events from Social Media. In *Proc. of ICWSM*, 2013.

[6] M. Condorcet. Essay on the Application of Analysis to the Probability of Majority Decisions. 1785.

[7] J. M. Conroy, J. D. Schlesinger, and D. P. O'Leary. Nouveau-ROUGE: A Novelty Metric for Update Summarization. *Computational Linguistics*, 37(1), 2011.

[8] P. C. de Oliveira, E. W. Torrens, A. Cidral, S. Schossland, and E. Bittencourt. Evaluating Summaries Automatically – a System Proposal. In *Proc. of LREC*, 2008.

[9] G. Giannakopoulos and V. Karkaletsis. AutoSummENG and MeMog in Evaluating Guided Summaries. In *Proc. of TAC*, 2011.

[10] G. Giannakopoulos, V. Karkaletsis, G. Vouros, and P. Stamatopoulos. Summarization System Evaluation Revisited: N-gram Graphs. *ACM Transactions on Speech and Language Processing*, 5(3), 2008.

[11] C.-Y. Lin. ROUGE: A Package for Automatic Evaluation of Summaries. In *Proc. of ACL*, 2004.

[12] C.-Y. Lin and E. Hovy. The Automated Acquisition of Topic Signatures for Text Summarization. In *Proc. of COLING*, 2000.

[13] C.-Y. Lin and E. Hovy. Automatic Evaluation of Summaries using N-gram Co-occurrence Statistics. In *Proc. of NAACL*, 2003.

[14] J. Lin. Divergence Measures based on the Shannon Entropy. *IEEE Transactions on Information Theory*, 37(1), 1991.

[15] E. Lloret and M. Palomar. Compendium: a Text Summarisation Tool for Generating Summaries of Multiple Purposes, Domains, and Genres. *Natural Language Engineering*, 19(2), 2013.

[16] A. Louis and A. Nenkova. Automatically Assessing Machine Summary Content without a Gold Standard. *Computational Linguistics*, 39(2), 2013.

[17] C. Macdonald. *The Voting Model for People Search. PhD Thesis*, University of Glasgow, 2009.

[18] R. McCreadie, I. Soboroff, J. Lin, C. Macdonald, I. Ounis, and D. McCullough. On Building a Reusable Twitter Corpus. In *Proc. of SIGIR*, 2012.

[19] S. Mithun, L. Kosseim, and P. Perera. Discrepancy Between Automatic and Manual Evaluation of Summaries. In *Proc. of WEAS*, 2012.

[20] A. Nenkova and K. McKeown. Automatic Summarization. *Foundations and Trends in Information Retrieval*, 5(2-3), 2011.

[21] A. Nenkova and R. Passonneau. Evaluating Content Selection in Summarization: The Pyramid Method. In *Proc. of HLT-NAACL*, 2004.

[22] A. Nenkova, R. Passonneau, and K. McKeown. The Pyramid Method: Incorporating Human Content Selection Variation in Summarization Evaluation. *ACM Transactions on Speech and Language Processing*, 4(2), 2007.

[23] A. Nenkova and L. Vanderwende. The Impact of Frequency on Summarization. *Tech. Rep., Microsoft Research, MSR-TR-2005-101*, 2005.

[24] J. Nichols, J. Mahmud, and C. Drews. Summarizing Sporting Events using Twitter. In *Proc. of IUI*, 2012.

[25] K. Owczarzak, J. M. Conroy, H. T. Dang, and A. Nenkova. An Assessment of the Accuracy of Automatic Evaluation in Summarization. In *Proc. of WEAS*, 2012.

[26] P. A. Rankel, J. M. Conroy, H. T. Dang, and A. Nenkova. A Decade of Automatic Content Evaluation of News Summaries: Reassessing the State of the Art. In *Proc. of ACL*, 2013.

[27] K. D. Rosa, R. Shah, B. Lin, A. Gershman, and R. Frederking. Topical Clustering of Tweets. *Proc. of SIGIR*, 2011.

[28] D. Rout, K. Bontcheva and M. Hepple. Reliably Evaluating Summaries of Twitter Timelines. *Proc. of AAAI-AMW*, 2013.

[29] Y. Sankarasubramaniam, K. Ramanathan, and S. Ghosh. Text Summarization using Wikipedia. *Information Processing & Management*, 50(3), 2014.

[30] B. P. Sharifi, D. I. Inouye, and J. K. Kalita. Summarization of Twitter Microblogs. *The Computer Journal*, 109, 2013.

[31] A. Singhal, C. Buckley, and M. Mitra. Pivoted Document Length Normalization. In *Proc. of SIGIR*, 1996.

[32] K. Spärck Jones. Automatic Summarizing: Factors and Directions. *Advances in Automatic Text Summarization*, 1999.

[33] K. Spärck Jones. Automatic Summarising: The State of the Art. *Information Processing & Management*, 43(6), 2007.

[34] T. Tideman. Independence of Clones as a Criterion for Voting Rules. *Social Choice and Welfare*, 4(3), 1987.

[35] A. Zubiaga, D. Spina, E. Amigó, and J. Gonzalo. Towards Real-time Summarization of Scheduled Events from Twitter Streams. In *Proc. of HT*, 2012.