FP7-ICT Strategic Targeted Research Project (STREP) TrendMiner (No. 287863)

Large-scale, Cross-lingual Trend Mining and Summarisation of Real-time Media Streams

D4.3.1 Evaluation results report

**Dominic Rout (The University of Sheffield), Ian Roberts (The University of Sheffield),
Kalina Bontcheva (The University of Sheffield)**

Abstract.
FP7-ICT Strategic Targeted Research Project (STREP) ICT-2011-287863 TrendMiner
Deliverable D4.3.1 (WP4.3.1)

This deliverable presents the evaluation of the TrendMiner tweet summarisation systems using a gold standard data set. The results show that the best approaches considerably outperform strong baselines, using the MAP and ROUGE-1 evaluation metrics. We carried out also a limited user-based study, which unfortunately could not verify the results from the automatic evaluation experiments. Following a discussion with the political use case partner involved in the user-based study, we concluded that either the requirements of the political use case itself, or the criteria used to identify tweets that meet those user needs, may have changed since the gold standard used for automatic evaluation was created initially.

D4.3.1 also evaluates the performance of the multi-paradigm indexing tool described in deliverable 4.2.1, to inform our design of indexing configurations that are able to sustain throughput of millions of Tweets per hour, as required for (near) real-time tweet analysis.

Keyword list: evaluation, gold standard, ROUGE, MAP, summarisation, tweet ranking, twitter, user study

Project	TrendMiner No. 287863
Delivery Date	October 30, 2014
Contractual Date	October 30, 2014
Nature	Report
Reviewed By	Thierry Declerck (DFKI)
Web links	NA
Dissemination	PU

TrendMiner Consortium

This document is part of the TrendMiner research project (No. 287863), partially funded by the FP7-ICT Programme.

DFKI GmbH
Language Technology Lab
Stuhlsatzenhausweg 3
D-66123 Saarbrcken
Germany
Contact person: Thierry Declerck
E-mail: declerck@dfki.de

University of Southampton
Southampton SO17 1BJ
UK
Contact person: Mahesan Niranjan
E-mail: mn@ecs.soton.ac.uk

Internet Memory Research
45 ter rue de la Révolution
F-93100 Montreuil
France
Contact person: France Lafarges
E-mail: contact@internetmemory.org

Eurokleis S.R.L.
Via Giorgio Baglivi, 3
Roma RM
00161 Italy
Contact person: Francesco Bellini
E-mail: info@eurokleis.com

Institute of Computer Science
Polish Academy of Sciences
5 Jana Kazimierza Str
01-248 Warsaw
Poland
Contact person: Maciej Ogrodniczuk
E-mail: Maciej.Ogrodniczuk@ipipan.waw.pl

Universidad Carlos III de Madrid
Av. Universidad, 30
28911 Madrid
Spain
Contact person: Paloma Martínez Fernández
E-mail: pmf@inf.uc3m.es

University of Sheffield
Department of Computer Science
Regent Court, 211 Portobello St.
Sheffield S1 4DP
UK
Contact person: Kalina Bontcheva
E-mail: K.Bontcheva@dcs.shef.ac.uk

Ontotext AD
Polygraphia Office Center fl.4,
47A Tsarigradsko Shosse,
Sofia 1504, Bulgaria
Contact person: Atanas Kiryakov
E-mail: naso@sirma.bg

Sora Ogris and Hofinger GmbH
Bennogasse 8/2/16
A-1080 Wien
Austria
Contact person: Christoph Hofinger
E-mail: ch@sora.at

Hardik Fintrade Pvt Ltd.
227, Shree Ram Cloth Market,
Opposite Manilal Mansion,
Revdi Bazar, Ahmedabad 380002
India
Contact person: Suresh Aswani
E-mail: m.aswani@hardikgroup.com

DAEDALUS - DATA, DECISIONS
AND LANGUAGE, S. A.
C/ López de Hoyos 15, 3
28006 Madrid
Spain
Contact person: José Luis Martínez Fernández
E-mail: jmartinez@daedalus.es

Research Institute for Linguistics of the
Hungarian Academy of Sciences
Benczr u. 33,
1068 Budapest Hungary
Contact person: Tamás Váradi
E-mail: varadi.tamas@nytud.mta.hu

Executive Summary

A number of tweet summarisation techniques have been proposed for the exploration of content related to the task of socio-political analysis and exploration. We focus on re-ranking whole tweets (as opposed to textual summarisation), and have developed, in collaboration with partners in the domain, a gold standard for tweet recommendation, where tweets are first selected according using named entities from a comparitively large background corpus, and then annotated for relevance.

Our tweet ranking gold standard will be made publically available, and has been used to automatically evaluate the performance of a number of tweet rankers, including a selection of approaches from the related work and strong baseline methods. We found, using ROUGE and MAP, that applying dimensionality reduction with topic models prior to the centroid algorithm could significantly outperform all other evaluated approaches, for our data set.

In order to evaluate further and more rigorously, we designed and carried out an offline user study, showing the output from our best approaches to between two and three annotators, who each gave subjective judgements on them in terms of utility, redundancy, completeness and preference. On the 17 documents and collections of 5 corresponding summaries, it was found that a random baseline was preferred on average above the best performing centroi ranker, though we could not state the significance of the preference.

It is possible that the information needs of the partner that generated the gold standard have changed over time. We partially verified this by allowing them to reannotate some tweet sets from the original gold standard, finding only weak agreement between the two sets of judgements by the same annotators. More qualitative research may be needed to discover the source of these discrepencies.

In addition, this deliverable describes an evaluation of the performance characteristics of the multi-paradigm indexing system described in deliverable 4.2.1, which determined that the system can sustain indexing rates of millions of Tweets per hour when configured appropriately. If a single index is insufficient for a given data rate the system can be transparently scaled by adding more indexes working in parallel, but presented as a single federated index to clients.

Contents

Chapter 1

Introduction

In deliverable 4.1.2 the problem of information overload for tweets was outlined and several candidate systems were developed. D4.1.2 also reported on some prelimary automatic evaluation of these methods, reporting simple Mean Average Precision (MAP) results for a gold standard data set produced by SORA.

In this deliverable, we expand the evaluation of the proposed summarisation systems considerably, reporting the results using more principled evaluation metrics, investigating powerful baselines and performing a small scale user-based evaluation experiment. We argue in Section 2.1 that the MAP evaluation metric is very limited, when applied to summary evaluation and address this weakness by investigating the ROUGE summary evaluation metrics.

We additionally attempt to improve the reliability of our results by corroborating judgements produced by SORA with those created by other German speakers as part of a user-based evaluation experiment. As such, this deliverable repeats some of the automatic evaluation results, augmented with new scores using the ROUGE metric. In some cases, the algorithms have been updated too and the newest results are given.

We do not address comparison to MEAD or SUMMA, which are popular textual summarisation frameworks. Many of the algorithms contained in these frameworks (cosine similarity, centroid, etc.) are sufficiently common and well understood, that we were able to reproduce them within our own Tweet ranking framework. Besides, MEAD and SUMMA have been developed for summarisation of longer, coherent documents and their application to the noisy, multi-topic tweet timelines, without any genre adaptation, is likely to produce not very good results. Instead, we compare against other state-of-the-art tweet recommendation algorithms.

Within the context of the project, political analysts from SORA were consulted on their media monitoring practices and Twitter tools used. Based on their requirements to go beyond simple keyword, hashtag, and user mention tweet searches, a temporally-aware interactive summarisation approach was designed:

1. A user chooses a temporally bounded window of tweets from a set of available time periods (typically hourly intervals).
2. Based on this selection, topics and named entities are extracted automatically and visualised in an interactive word cloud.
3. The user then chooses one or more topics and entities of interest, and thus narrows down the set of tweets to be explored.
4. The resulting tweets are ranked for relevance and only the top n tweets are displayed.

This process is described in greater detail in Deliverable 4.1.2. The evaluation described in this deliverable primarily measures the quality of automated systems to address the fourth step in the above sequence.

This report also includes details of some experiments to assess the performance of the multi-paradigm indexing system described in Deliverable 4.2.1. When used with streaming data sources where new data is arriving continuously, it is important to be sure that the tools are able to process data as fast as it arrives. These experiments provide some insight into the optimal configuration of the indexing system and the necessary hardware to handle a given number of Tweets per hour.

1.1 Relevance to TrendMiner

Since TrendMiner is targetted at keyword filtered and time windowed views from large scale streaming media, summarisation is necessary where such collections are not feasibly consumed manually. Since it is important to prioritise useful and interesting content to assist analysis, we must manually evaluate algorithms that attempt to provide this functionality. A user-focussed evaluation allows us to accurately demonstrate the extent to which our suggested approaches are useful in the context of the TrendMiner use cases.

1.1.1 Relevance to project objectives

This deliverable presents an in depth evaluation of the methods for semantic search and summarisation of streaming social media content.

1.1.2 Relation to other workpackages

The summaries generated in this work package are integrated as part of a larger TrendMiner system in WP5. The tweet filtering part of the annotation task is carried out using named entities disambiguated by the system developed as part of WP2.

Chapter 2

Automatic evaluation of Tweet Summaries

This chapter includes some results originally reported in Deliverable 4.1.2. However we have updated some of the algorithms in question and thus report updated results for these. This deliverable also reports results using ROUGE, an evaluation metric that is better suited to text summarisation and which gives a fuller picture of performance.

Deliverable 4.1.2 also described the creation of a gold standard for tweet relevance. The gold standard is comprised of tweet sets, which have been filtered according to time and keywords, and binary judgements indicating which tweets from each are considered most interesting by annotators from SORA. This gold standard is available publically at `https://gate.ac.uk/data/trendminer/tweet_relevance_de.html`. The wording of this task is shown in Sections A.1.1 and A.1.2

This chapter presents the comparative evaluation results on this gold standard for a number of summarisation approaches. Such automatic evaluation has several distinct advantages. The human effort of creating the gold standard is expended only once, and it may be quickly applied to many algorithms. Additionally, it is straightfoward to perform significance analysis for results produced in this way.

The fully automated evaluation approach has many limitations, which will be discussed further in 3. Importantly, one must use a suitable evaluation metric, to compare the ideal manual summaries with the candidate automatic ones. This can be achieved using a measure of textual similarity, or manually by comparing key points in the Pyramid method [NP04]. The Pyramid method is arguably the most complete way to compare a candidate summary with one in a gold standard, but as it requires human annotators, it is no longer fully automated and therefore cannot be used to evaluate a large number of systems and parameters.

2.1 Scoring Summaries

The judgements produced by the gold standard annotators were binary, though the systems we evaluate are intended to produce a continuous ranking of tweets. For this reason, Mean Average Precision (MAP) [MRS08] is used to form a simple impression of performance where the number of tweets read by a user is allowed to vary.

Given $P(k)$ as fraction of documents that are relevant out of those appearing before position k in a ranking, and $rel(k)$, an indicator function, is 1 if the document in position k is relevant and 0 otherwise, we can write average precision for a specific query as:

$$\text{Average Precision} = \frac{\sum_{k=1}^{n}(P(k) \times \text{rel}(k))}{\text{\#Relevant Documents}} \tag{2.1}$$

When our hypothetical user reaches an interesting post, the precision is the ratio of interesting over uninteresting posts they would have had to read to get there. The same calculation is carried out for all relevant posts in a set, and averaged to give Average Precision. The mean of these averages across all sets is the MAP.

This measure however has some disadvantages:

- It tests only whether the exact same post is included, ignoring textual similarity to the gold standard.
- Likewise, where several very similar posts exist in a collection, a reasonable user may select only one. A ranking system which considers diversity may well select another and be incorrectly penalised.
- Average precision can vary greatly from query to query.

Therefore, in addition to MAP, we also report scores according to the ROUGE-1 metric, which is used in text summarisation[Lin04]. The ROUGE-1 version of ROUGE considers unigrams only, and terms from a german stop word list are excluded[Por01]. ROUGE-1 consists of three measures which are recall, precision, and a combined F-score. As ROUGE is not specifically designed for a ranking task, the top 8 ranked tweets are used to form a candidate summary.

2.2 Results

2.2.1 Baselines

Table 2.1 shows the results of several $s$ocial feature baselines individually, e.g. ranking tweets based on the number of retweets or favourites they have.

The results show that the number of times a tweet has been favourited by users is the best performing social baseline. On the other hand, retweets are significantly weaker.

Temporal ranking alone also performs very poorly. We showed the tweets in a random order when creating the gold standard, demonstrating that although Twitter displays posts reverse-chronologically, there is little implicit preference for newer tweets. In other words, the current ordering used in the Twitter web interface is sub-optimal for users who prefer tweets ordered by relevance.

The number of lists on which a user has been placed performs considerably worse than all of other social baselines. It is possible that this feature is simply too coarse to be useful for ranking on its own. We did not attempt in this work to combine twitter-specific features with text-based features, as we do not have a large-enough training corpus. In addition, such methods tend to rely not only on the social relationships of the author, but also on their connection to the reader [YLL12, UC11] whereas in our dataset and media monitoring task we do not have information about the specific reader.

Algorithm	MAP	ROUGE-R	ROUGE-P	ROUGE-F
Number of Retweets	23.44%	31.79%	28.01%	29.53%
Number of Favourites	**27.58%**	**36.29%**	**36.88%**	**36.28%**
Number of lists containing author	22.02%	26.80%	27.96%	27.11%
Time created	19.99%	22.89%	22.93%	22.71%

Table 2.1: Performance for social network baselines

Several textual baselines were evaluated (see Table 2.2). Cosine similarity outperforms most of the social baselines (with the exception of favourites), though using IDF weighting significantly worsens these scores. This is most likely due to the small size of the vocabulary being considered, which focussed around political topics, and the fact that queries were generated using selection from a list of name entities, therefore containing no stop words. Hybrid TF.IDF [IK11] exploits word counts at multiple scopes and includes a threshold on redundancy. It apparently outperforms cosine similarity, but not the favourites baseline. We set the threshold (0.85) for Hybrid TF.IDF on this dataset using grid search.

Lastly, we hypothesized that random reordering should be a weak baseline, due to the fixed size of the candidate set and the fixed number of positive judgements. In such cases, the score for random reordering under MAP is simply a random sample related to that metric. The score for random reordering under ROUGE is higher, but is determined by the textual variance of documents within the same set; since the documents are pre-filtered by the user search, this variance is low and ROUGE is higher than MAP for the random case.

Algorithm	MAP	ROUGE-R	ROUGE-P	ROUGE-F
Hybrid TF.IDF	28.65%	30.92%	**39.94%**	**34.50%**
Cosine	**30.20%**	**32.45%**	35.38%	33.59%
Cosine with IDF	25.99%	26.68%	31.01%	28.42%
Random reordering	22.75%	27.55%	28.40%	27.69%

Table 2.2: Performance for information retrieval

2.2.2 Centroid and TextRank

The second set of experiments evaluated the Centroid [RJST04] and TextRank [MT04] algorithms, with a number of textual n-gram features (see Table 2.3). In the case of unigrams, we also experimented with IDF weighting, derived from hourly sections of the whole timeline. For bigrams and trigrams no weighting was used as we did not have an extremely large training data set.

Even though the ROUGE results for TextRank are higher than those for Centroid, the scores did not differ significantly (p=0.33 for unweighted unigrams).

Algorithms that used unigrams with IDF appeared more effective those that used unweighted terms, though the differences in ROUGE were also not significant (p=0.72 for Centroid unigram, p=0.06 for TextRank unigram). Likewise, we did not find significant differences when including bigrams alonside weighted unigrams (p=0.8 for centroid, p=0.25 for TextRank.

Although some of the textual approaches appear to outperform the baselines using social features as described in section 2.2.1, the best performing, TextRank with unigrams and bigrams, is not significantly stronger than the counts of favourites baseline (P=0.58).

2.2.3 Dimensionality Reduction Results

The next experiment tested the hypothesis that dimensionality reduction methods help with tweet ranking, when compared to n-gram based features. In particular, topic models are used to map the higher dimensional n-gram feature space onto a lower dimensional space of topics. We then proceed to use Centroid as before, in order to rank tweets. The topic models are created using Latent Semantic Indexing (LSI) [DDF$^+$90] and Latent Dirichlet Allocation (LDA) [BNJ03].

As shown in Figure 2.1, using topic models in this way significantly outperforms using Centroid ranking with IDF alone (p=0.008, LSI 200 topics), as well as the favourites baseline (p=0.017, LSI 200 topics). The reported figures are for unigrams with IDF. We do not incorporate LSI or LDA as a preprocessing component for TextRank, as the resulting graphs are generally very dense whereas TextRank performs well on sparser text graphs.

The best performance was achieved by LSI with 200 topics, and using LDA performed consistently worse than LSI. However, due to the limited size of our dataset, it is not pos-

Algorithm	Features	MAP	ROUGE-R	ROUGE-P	ROUGE-F
Centroid	Unigram	26.61%	33.73%	31.30%	32.26%
	Unigram (case preserved)	25.93%	32.82%	30.14%	31.22%
	Unigram with IDF	33.04%	36.36%	35.66%	35.76%
	Unigram with IDF (case preserved)	32.00%	35.42%	34.45%	34.67%
	Bigram only	30.67%	34.16%	33.47%	33.44%
	Unigram with IDF & bigram	33.27%	36.49%	**35.97%**	**35.98%**
	Trigram	29.57%	35.45%	31.25%	32.98%
	Unigram with IDF, bigram & trigram	**33.30%**	**36.59%**	35.49%	35.82%
TextRank	Unigram	27.45%	33.89%	32.89%	33.17%
	Unigram (case preserved)	26.70%	33.14%	31.63%	32.16%
	Unigram with IDF	32.15%	35.69%	40.01%	37.43%
	Unigram with IDF (case preserved)	30.86%	34.79%	38.62%	36.36%
	Bigram only	27.30%	30.54%	34.47%	32.01%
	Unigram with IDF & bigram	33.60%	**35.92%**	**40.51%**	**37.76%**
	Trigram	27.64%	32.12%	33.78%	32.72%
	Unigram with IDF, bigram & trigram	**33.78%**	35.83%	40.18%	37.58%

Table 2.3: Performance of Centroid and TextRank

sible to draw a reliable conclusion that LSI is always better than LDA for tweet ranking. What we can conclude, however, is that considerable performance improvement can be seen when using topic-based dimensionality reduction methods.

2.2.4 Maximal Marginal Relevance

For our experiments we use both TextRank with IDF and bigrams, and our LSI topic models, which were our best performing approaches without and with dimensionality reduction respectively. Diversity is incorporated using the Maximal Marginal Relevance (MMR) algorithm [CG98]. The same feature representation was used when calculating redundancy as for the centroid. Where centroid scores are calculated using feature vectors that result from dimensionality reduction with LSI, the redundancy score is calculated using cosine distance to tweets in the existing summary in that same lower-dimensional space.

Model	Topics	MAP	ROUGE-R	ROUGE-P	ROUGE-F
Latent Semantic Index	10 topics	33.50%	38.49%	36.90%	37.41%
	50 topics	33.83%	38.36%	38.43%	38.17%
	100 topics	39.68%	43.43%	44.14%	43.59%
	200 topics	**41.44%**	**44.54%**	**46.67%**	**45.36%**
	400 topics	40.49%	43.40%	45.40%	44.14%
	600 topics	39.51%	43.14%	45.44%	44.01%
	800 topics	38.37%	41.70%	44.13%	42.62%
	1000 topics	38.34%	42.34%	44.86%	43.30%
Latent Dirichlet Allocation	10 topics	25.74%	30.44%	30.66%	30.36%
	50 topics	28.63%	35.62%	37.87%	36.52%
	100 topics	29.71%	33.63%	34.97%	34.11%
	200 topics	27.85%	32.35%	34.64%	33.24%
	400 topics	34.40%	40.41%	42.94%	41.41%
	600 topics	**38.31%**	**41.47%**	**44.46%**	**42.72%**
	800 topics	38.24%	40.17%	43.65%	41.63%
	1000 topics	34.60%	37.70%	41.45%	39.27%

Table 2.4: Performance with Dimensionality Reduction

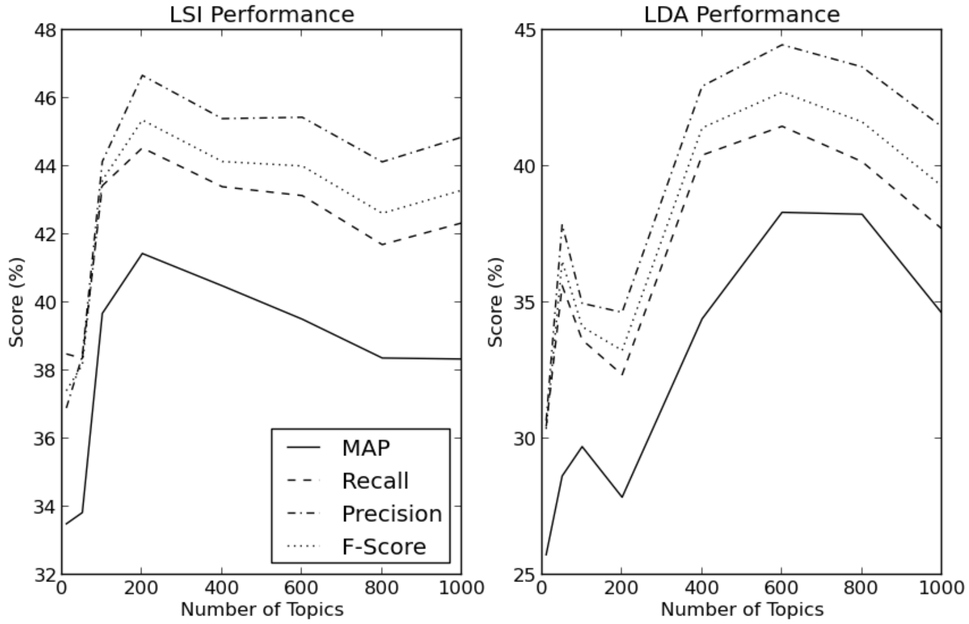


Figure 2.1: Performance with Dimensionality Reduction

A λ of 0.0 is maximal diversity, whereas 1.0 is equivalent to centroid similarity.

Features	λ	MAP	ROUGE-R	ROUGE-P	ROUGE-F
TextRank	0.0	18.76%	21.77%	23.80%	22.51%
	0.2	21.10%	25.67%	26.96%	26.11%
	0.4	21.18%	26.46%	27.79%	26.91%
	0.6	21.41%	26.36%	27.80%	26.84%
	0.8	23.16%	28.11%	29.34%	28.47%
	1.0	**33.60%**	**35.92%**	**40.51%**	**37.76%**
LSI (200 topics)	0.0	17.98%	20.66%	21.66%	20.91%
	0.2	20.07%	23.18%	23.81%	23.29%
	0.4	22.96%	28.63%	28.58%	28.36%
	0.6	28.55%	35.40%	35.32%	35.17%
	0.8	35.28%	41.82%	42.38%	41.87%
	1.0	**41.44%**	**44.54%**	**46.67%**	**45.36%**

Table 2.5: Performance with MMR

Our intention with the inclusion of MMR was to prevent the reranked tweets from placing many, very similar tweets at the top of the ranking, and to ensure that at a glance a user would be able to see all of the topics in contained in the set. Contrary to our intuitions, the performance when MMR is applied was found to be worse than without MMR, and the results worsened the more strongly we prioritised novelty. Note that in the worse cases the MAP for MMR is worse than that of random reordering, showing that MMR is specifically de-emphasising interesting posts.

Redundancy reduction is harmful rather than helpful in this case, which may suggest that the gold standard annotators were interested in finding canonical tweets which best represent the topic at hand, regardless of how much they repeat one another.

2.3 Discussion

Most of the evaluated tweet ranking methods outperformed significantly the reverse-chronological baseline. This demonstrates that current Twitter monitoring tools need to cater better for data exploration tasks, and, in particular, to implement better relevance ranking methods. In addition, we demonstrated that retweet counts are also not a good predictor of tweet relevance. A new finding of this research is that the number of times a tweet has been favourited by users forms a very strong baseline. This is outperformed only when dimensionality reduction is used on the textual features, to create topic-based features.

The significant performance improvements, when LDA and LSI are used for dimensionality reduction, indicate that dimensionality is a significant issue in tweet ranking for social media monitoring. It indicates that analysts may have an information need, and be able to articulate that through exploratory search, but they may not be able to express all of the relevant vocabularity, nor indicate which terms are more useful to them than others.

Chapter 3

Manual Evaluation of Tweet Summaries

The automatic evaluation described in chapter 2 is useful for comparing numerous algorithms and approaches without repeated expensive human annotation. Automatic evaluation is, however, only an approximate substitute for real evaluation of how humans would judge system performance. As such, it is necessary to complement automatic evaluation of many relevant approaches with manual evaluation, which is focused on user-based judgement of the results of the best performing algorithms.

By carrying out a user study focused on our best performing system (and several relevant baselines), we address some of the weaknesses of the automated approach used in the rest of this work. In our gold standard generation, we fixed the number of selected tweets to 50 (using random sampling to achieve this where needed) and the number of tweets in the summary to 8. No such limitation is applied in the user study.

The evaluation sets are already filtered through the selection of groups of related terms in a tag cloud interface. As such, many of the sets are likely to contain duplicate tweets (both verbatim re-tweets and modifying tweets, where the tweet author has reused content from the original tweet). Therefore, it is entirely possible that two semantically identical tweets exist in the same set, and that the annotator marks only one of them as relevant. We attempt to mitigate this problem using ROUGE and by treating tweets that are identical to interesting tweets as also interesting. However, by doing so we fail to address the issue of the generated summaries perhaps containing unacceptable amounts of redundancy.

While redundancy could be addressed to some extent by presenting the summaries using an interface that would collapse several retweets into one tweet, in practise much of the redundancy is not just retweets but also tweets that address very similar content, perhaps a quote from a news resource or a politician, and such a method could easily become an additional summarisation step in itself.

Prior to carrying out the user study, we asked SORA to generate a new selection of keyword groups for the political use case from historical data. These filters were then used, rather than those from the gold standard generation process, as we wished to similate a scenario in which information is being discovered for the first time. This would be unrealistic had the tweets already been considered in detail for a previous annotation task.

3.1 Study Design

We intend to compare the tweet rankings produced by five algorithms for *completeness*, *redundancy*, *utility* and *overall subjective preference*. These dimensions were chosen from work on evaluating recommender systems[SG11], which we argue is somewhat related to the task at hand. Our problem differs slightly from recommender systems in that we do not take preferences from one user to generate suggestions for another, and we do not personalise the recommendations.

The task at hand is one of discovering information, not one of recommending a product, and it is meant to be carried out on a small scale by known analysts, so we do not evaluate certain areas which might otherwise be evaluated for a recommender system, such as serendipity, robustness, cold start performance, scaleability, adaptability or privary.

Although the intention is the direct comparison between summaries, we do not ask annotators to rank the summaries (e.g. from best to worst) directly, as it could be confusing to perform this task for four different dimensions. We instead ask volunteers to score each criterion separately, using a Likert scale [Lik32], which differs in that users are allowed to assign the same scores to several summaries. Annotators are not specifically told to place the summaries in order of preference, though they may choose to do so.

The question used was as follows:

Please give your opinion on the following criteria, regarding the summary above:

- The summary captures all the important information from the full set of tweets (completeness)
- There were several tweets in the summary that were repeating very similar information (redundancy)
- I could use this summary to study political figures or events (utility)
- Given your responses above, please rate this summary as a whole (subjective preference)

All of the responses were given on a scale of 1 (Disagree) to 5 (Agree), apart from the final rating which was made from 1 to 7, with the responses ‘Very much dislike’, ‘Dislike’, ‘Somewhat dislike’, ‘Undecided’, ‘Somewhat like’, ‘Like’ and ‘Very much like’. The intuition behind the latter, larger Likert scale was to allow users to make finer grained distinctions for their overall summary preferences. The full interface used for this task is shown in Section A.2

Unlike in the gold standard creation task, we present the same tweet summaries to several annotators. This allows us to aggregate the results from several users with the intention of gaining a more robust evaluation result.

The particular experimental configuration used for the user study does not lend itself well to significance testing. The scores on the Likert scale are ordinal, but arguably not scalar, so differences cannot be reliably interpretted, unless one algorithm dominates the others, especially with so few users in the study.

To calculate agreement between respondents, we take the mean of pairwise spearman's Rho for each document set in the study. Spearman's Rho can be used to calculate correlation for ordinal data, as we have here[SG11]. To compare systems, we sort and rank them amongst each dimension according to the user response (to avoid differences in magnitude between users) and take the mean rank for that system across all filter sets.

Figure 3.1 shows an example summary and the interface used for evaluation. Several summaries were placed on the same page in a random order, and evaluated one after the other.

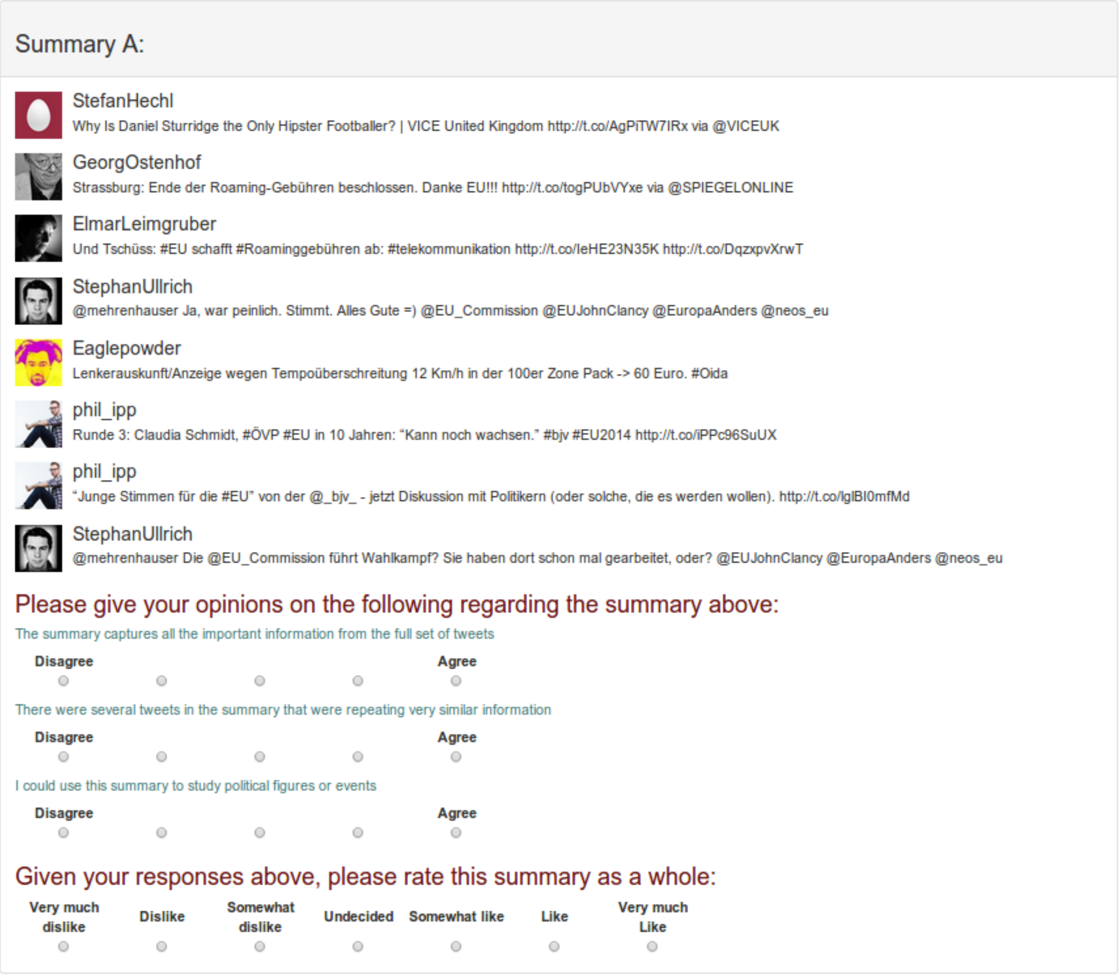
3.2 Evaluated approaches

Random This is an important baseline in which tweets are simply shuffled at random. We expect this to be relatively strong, as the tweets are to some extent already filtered according to the keywords they contain.

Retweets The count of the number of times each tweet has been retweeted is used to determine its ranking and thus its inclusion in the summary. Retweets can be considered strong evidence that at least some other Twitter users find the tweet interesting. We would expect a useful system to outperform retweets, though retweets are no always immediately available (for example shortly after a tweet is posted).

Hybrid TF.IDF We compare to hybrid TF.IDF, an approach to tweet ranking from related work[IK11]. Hybrid TF.IDF contains measures to help reduce redundancy in the generated summaries. Since the chosen automated evaluation does not always reward systems with greater diversity, hybrid TF.IDF may fare better within the user study than against the gold standard.

Phrase reinforcement Another system for comparison to related work[SHK10], phrase reinforcement attempts to leverge repeated word pairs in the text to generate textual summaries. Comparison with PR against a gold standard is extremely unreliable, since the summaries generated by this algorithm are short phrases rather than collections of tweets. Unfortunately, while users would not be able to map easily between a summary and the



Summary A:

StefanHechl
Why Is Daniel Sturridge the Only Hipster Footballer? | VICE United Kingdom http://t.co/AgPiTW7IRx via @VICEUK

GeorgOstenhof
Strassburg: Ende der Roaming-Gebühren beschlossen. Danke EU!!! http://t.co/togPUbVYxe via @SPIEGELONLINE

ElmarLeimgruber
Und Tschüss: #EU schafft #Roaminggebühren ab: #telekommunikation http://t.co/leHE23N35K http://t.co/DqzxpvXrwT

StephanUllrich
@mehrenhauser Ja, war peinlich. Stimmt. Alles Gute =) @EU_Commission @EUJohnClancy @EuropaAnders @neos_eu

Eaglepowder
Lenkerauskunft/Anzeige wegen Tempoüberschreitung 12 Km/h in der 100er Zone Pack -> 60 Euro. #Oida

phil_ipp
Runde 3: Claudia Schmidt, #ÖVP #EU in 10 Jahren: "Kann noch wachsen." #bjv #EU2014 http://t.co/iPPc96SuUX

phil_ipp
"Junge Stimmen für die #EU" von der @_bjv_ - jetzt Diskussion mit Politikern (oder solche, die es werden wollen). http://t.co/lglBl0mfMd

StephanUllrich
@mehrenhauser Die @EU_Commission führt Wahlkampf? Sie haben dort schon mal gearbeitet, oder? @EUJohnClancy @EuropaAnders @neos_eu

Please give your opinions on the following regarding the summary above:

The summary captures all the important information from the full set of tweets

Disagree				Agree
○	○	○	○	○

There were several tweets in the summary that were repeating very similar information

Disagree				Agree
○	○	○	○	○

I could use this summary to study political figures or events

Disagree				Agree
○	○	○	○	○

Given your responses above, please rate this summary as a whole:

Very much dislike	Dislike	Somewhat dislike	Undecided	Somewhat like	Like	Very much Like
○	○	○	○	○	○	○

Figure 3.1: Interface for manual summary evaluation

method used to produce it, this is not the case for summaries generated by the PR algorithm, as the short phrases easily stand out from collections of tweets. Nonetheless, we include phrase reinforcement as an example of a textual summarisation system which produces very terse results, allowing us to qualitatively understand if this is what the user preferred.

Centroid (with topic models) This was the best performing method we evaluated, according to the gold standard. Here we use LSI with 50 topics as a preprocessing step, and do not attempt to reduce redundancy.

3.3 Annotators

Although the political use case partners (SORA) know their domain and information needs better than anyone else, they were only able to provide a single annotator for this task. Therefore, we used two additional German speaking users from the University of Sheffield, who have not been involved previously in the project. Both users were German, not Austrian, and worked in computer science rather than political science, so their inclusion is more useful in evaluating the summarisers themselves, rather than its applicability to the use case.

3.4 Agreement

Inter-annotator agreement was measured using Spearman's rho, as the ratings are ordered categories and not intervals. These results were calculated using a set of 17 documents, for which 5 summaries each were evaluated. Standard deviation was calculated as the mean of the standard deviation per document. The mean correlation in each dimension was moderately positive, with a small standard deviation.

The full results for agreement are shown in Table 3.4. This agreement is encouraging and shows that the task at hand could be replicated with some success by annotators outside of SORA.

Dimension	Spearman's Rho	σ
Completeness	0.46	0.022
Rating	0.45	0.032
Redundancy	0.45	0.036
Utility	0.45	0.016

Table 3.1: Agreement between annotators for manual evalution

3.5 Results

The results for manual evaluation are shown in table 3.5. We show the mean rank for each dimension. Unfortunately, the random baseline appears to outperform the best performing automated tweet ranking method on all dimensions, and it is the strongest of all the approaches we evaluate on all considered areas apart from redundancy, where it came second.

Though the random ordering appears to outperform our best approach, it is worth noting that because the results are not scalar it is difficult to identify whether such differences are statistically significant, unless one result overwhelms or is overwhelmed by the others (as is the case with Phrase reinforcement). Nonetheless, it appears that Random is a very strong baseline, because the tweet sets are already somewhat filtered. Additionally, a post-evaluation discussion with the SORA annotator revealed that they felt redundancy was a problem in the summaries that they saw, and that this affected their rankings amongst all dimensions.

This preference for more diverse summaries with lower redundancy is not shown in the automatic evaluation, where methods which attempt to reduce redundancy perform worse than those which simply use centroid. There are many possible reasons why this difference may have occurred, including the change in experimental design, whereby annotators now see a complete summary rather than individual tweets from which they must select, and the passing of time meaning that less current context is held implicitly in the memories of the annotators. It is also possible that the the information needs of the SORA users have changed since they created the gold standard used for automatic evaluation.

The phrase reinforcement algorithm also performed very poorly on all but redundancy, where it may have succeeded because of the wording of the question (contains redundant information). The summaries it generated were very short, usually only a few words, and annotators claimed they were nonsensical. It is possible that the tweet sets that we provided for this task were too small and contained too little repetition for phrase reinforcement to work optimally.

Algorithm	Completeness	Rating	Redundancy	Utility
Phrase Reinforcement	0.49	0.41	**0.76**	0.30
Centroid (LSI 50 topics)	2.00	1.86	2.81	2.24
Retweet counts	2.27	2.19	2.38	2.16
Hybrid TF.IDF	2.59	2.59	2.22	2.62
Random	**2.65**	**2.95**	1.84	**2.68**

Table 3.2: Mean rank of 5 systems under manual evaluation

3.6 Revisiting automatic evaluation

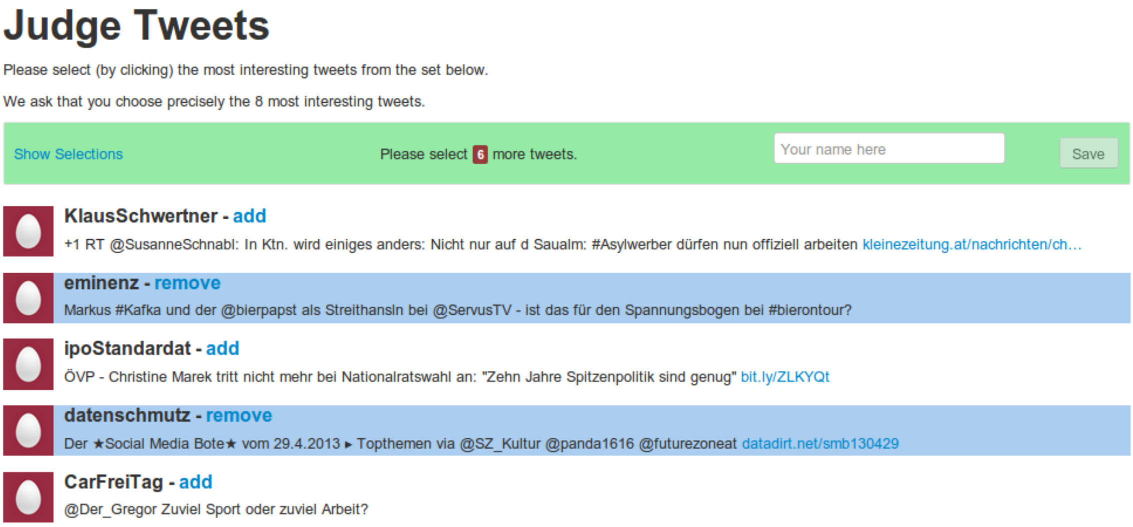
As discussed above, one of the hypotheses regarding the discrepancy between the results of the automatic evaluation and the user-based study, is that the information needs of SORA may have changed since the gold data corpus was first annotated. This would explain the difference in performance between the centroid method on automatic evaluation, in comparison to manual evaluation. In order to rule out this hypothesis, we requested SORA to re-annotate parts of the same gold standard data, without disclosing to them that they have annotated already these tweets in the past. This allowed us to measure agreement between the same SORA annotator at two different points in time. The first set of gold data annotations were created some time in October 2013, while the second set – on the 1st October 2014, roughly one year apart.

The annotation interface was slightly revised from the one used previously, to give less guidance (since the task had been carried out before) and with the intention of making the decisions more open ended. The revised annotation user interface is shown in Figure 3.2. While the basic request remained the same, we removed detail about why a tweet might be considered interesting. We hid the label tags and dates for each tweet set from the user, so as not to inadvertently cue to them that these were sets they had annotated before.

The observed agreement on a random selection of 15 tweet sets from the original gold standard was 0.0743 (Fleiss' Kappa), indicating only slight agreement. This result apparently demonstrates that the information needs of SORA have changed in the intervening time, or perhaps that memory and the currentness of the tweets in the collection play a role in the decisions made by annotators, somewhat confounding the reliance on automatic evaluation during development.

It is important also to note the weaknesses of this secondary task. While the annotations have indeed changed, one common comment from SORA throughout this work is that the sets generated by the keyword selection approach in particular contain many closely related tweets. Where the user has no particular preference between them, they may select differently at random, both times. Moreover, this secondary annotation demonstrated problematic elements of the original user study, in which ideally more users should have been involved, who should have been asked to revisit historical data after some time.

While it would be interesting to determine what effect if any this difference between SORA in 2013 and in 2014 has on the performance of our systems as evaluated with ROUGE, in practise the size of the set is very small and it would have been very resource intensive at a late stage of the project to revisit enough sets to allow robust comparison between algorithms. Nevertheless we intend to carry out analysis to determine significance of the ROUGE and MAP scores on the newer set and whether these differ demonstrably between those and the older set.



Judge Tweets

Please select (by clicking) the most interesting tweets from the set below.

We ask that you choose precisely the 8 most interesting tweets.

KlausSchwertner - add

+1 RT @SusanneSchnabl: In Ktn. wird einiges anders: Nicht nur auf d Saualm: #Asylwerber dürfen nun offiziell arbeiten kleinezeitung.at/nachrichten/ch...

eminenz - remove

Markus #Kafka und der @bierpapst als Streithansln bei @ServusTV - ist das für den Spannungsbogen bei #bierontour?

ipoStandardat - add

ÖVP - Christine Marek tritt nicht mehr bei Nationalratswahl an: "Zehn Jahre Spitzenpolitik sind genug" bit.ly/ZLKYQt

datenschmutz - remove

Der ★Social Media Bote★ vom 29.4.2013 ▸ Topthemen via @SZ_Kultur @panda1616 @futurezoneat datadirt.net/smb130429

CarFreiTag - add

@Der_Gregor Zuviel Sport oder zuviel Arbeit?

Figure 3.2: Revised manual annotation interface

Chapter 4

Indexing Performance for the Multi-Paradigm Search System

Deliverable 4.2.1 described the Mímir multi-paradigm indexing and search tools used in TrendMiner to store and provide access to the results of ontology-based information extraction (OBIE) over streaming media. In this deliverable we present statistics on the indexing speed that can be achieved by these tools, as a means of verifying that the tools will in fact scale to the quantities of data that need to be indexed in real-time, and to gain an insight into the hardware requirements and optimal software configurations for a production system.

The evaluation was based on a data set of 5.3 million Tweets that have previously been annotated using the TrendMiner OBIE system LODIE. Test harness software submitted these annotated Tweets to the indexing server and recorded the number of Tweets submitted and the time taken, which could be converted into a measurement of *Tweets indexed per hour (tph)*. The experiment was repeated with various permutations of indexing configuration and different numbers of submission threads running in parallel, with the cumulative indexing rate measured every 1,000 submitted Tweets.

To avoid artificial limitations imposed by network bandwidth, both the Mímir indexing system and the submission test harness were run on the same machine, a 20-core Ubuntu server with 256GB of RAM. The server offered both conventional hard disk storage and high speed solid-state storage, and experiments were tried with indexes stored on both types of storage, but this made little difference to the speed, and only the SSD measurements are reported here.

4.1 Initial experiments

The test data set was annotated with:

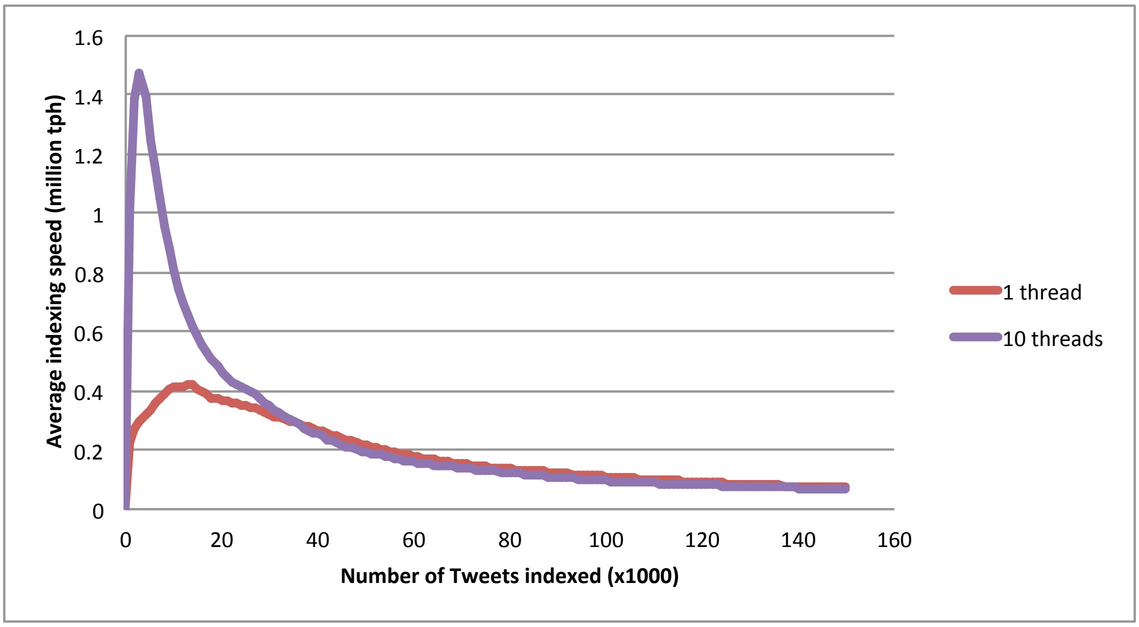


Figure 4.1: Initial experiment, including indexing of Tweet IDs

- "Token" annotations, i.e. the individual words within the Tweet, including their part of speech (POS) tag.
- "Mention" annotations, named entities including an instance URI within the DBpedia knowledge base.
- Tweet-level metadata: the Tweet ID and the user ID of the Tweet's author.

For a first experiment we attempted to index all this information in Mímir, using first a single submission thread, and then repeating the test with ten parallel threads. Figure 4.1 shows the results, measured as the average speed achieved up to a given point, i.e. the X axis denotes the number of Tweets n indexed so far and the Y axis is the indexing speed calculated from the cumulative time t taken by the indexing process up to this point, as:

$$\text{tph} = 3600\frac{t}{n}$$

It can be seen from the graph that indexing is initially very fast, with the ten thread case peaking at almost 1.5 million tph, but that as more and more Tweets are added, the rate becomes progressively slower. The reason for this is related to the way Mímir encodes annotation information for indexing. The underlying search engine, MG4J, operates on plain string terms, so Mímir must map semantic annotations to identifiers that can be handled as terms by MG4J. To optimise performance at search time, the approach used is to treat all annotations that share the same combination of annotation type, feature values and length (number of tokens spanned) as the same term. Therefore, when presented with

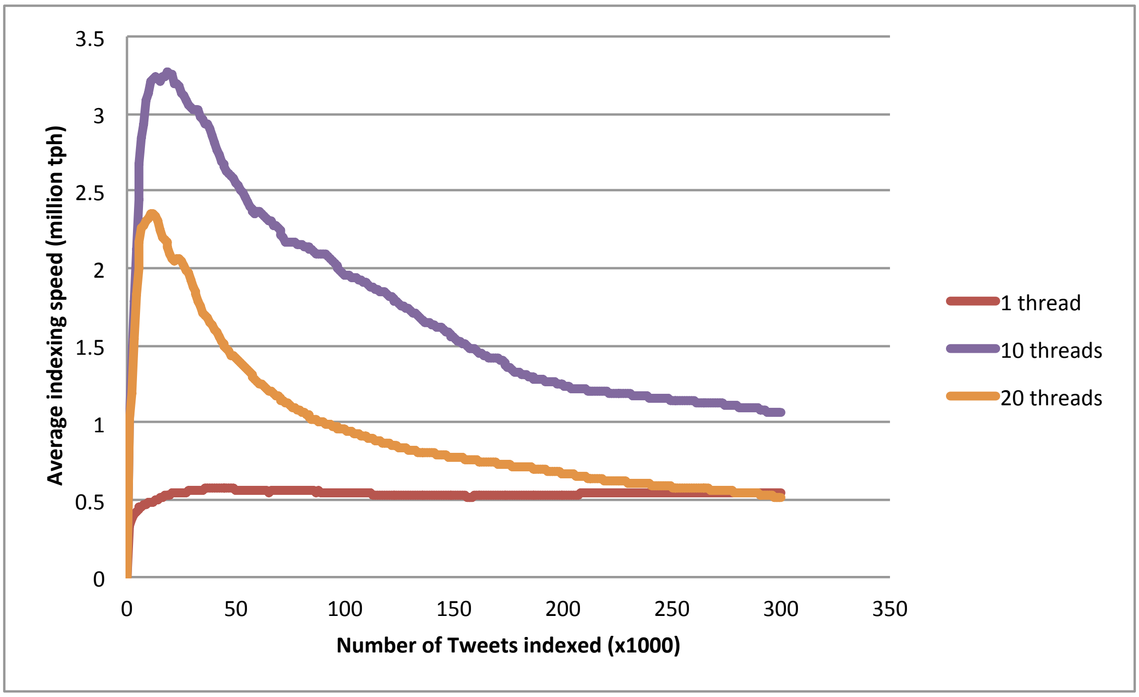


Figure 4.2: Testing different numbers of submission threads

a new annotation for indexing, Mímir must determine whether another annotation with the same combination of feature values has been seen before, and either extract the existing term string or assign a new one, as appropriate.

Unfortunately, this approach leads to certain pathological cases at indexing time, most notably when dealing with annotations (such as the Tweet ID) where *every* annotation has a unique combination of feature values. In this situation, every new annotation requires the insertion of a new row into the database mapping tables.

4.2 Excluding the Tweet ID

To counter this, a second set of experiments was run, where the Tweet ID feature was excluded from indexing. The results are shown in figure 4.2, and while the graphs show the same overall shape as before, the actual numbers are much better. The test was run with various numbers of threads submitting Tweets to the index, and here we see that ten threads provided the best performance. A single thread was unable to supply Tweets fast enough to keep the indexing server busy, and 20 threads caused too much contention.

As before, the rate peaks early on and then gradually declines as the number of distinct annotations increases. The rate of decline will depend on the exact combination of annotations found in the source data, but given a closed (albeit large) set of possible different feature values, we would expect the indexing speed to eventually level off once the

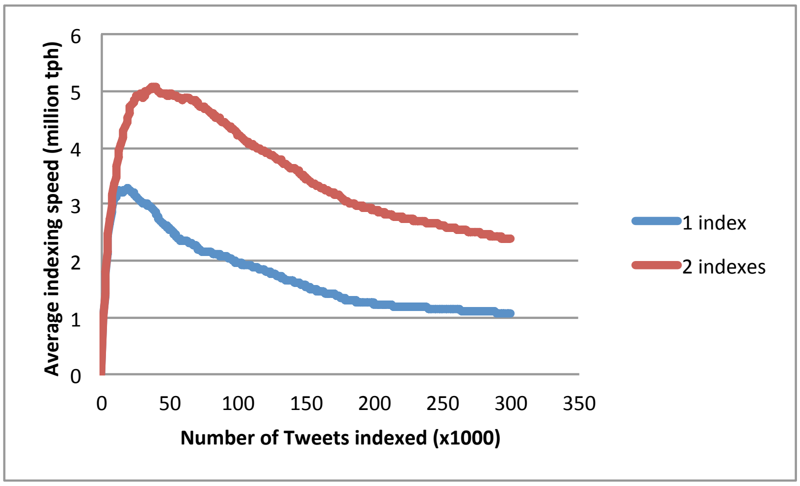


Figure 4.3: Using a federated index

majority of distinct entities have been seen at least once.

4.3 Scaling using multiple indexes

The Mímir indexing process is essentially a sequential operation. A single index consists of several sub-indexes which can be built in parallel, but submitted documents must be "pipelined" to ensure they are consumed by all sub-indexes in the same order (while sub-index 2 is processing document 1, sub-index 1 can start processing document 2, etc.). Therefore the maximum number of processor cores that can be occupied building a single index is equal to the number of sub-indexes. To utilise more processing capacity (on the same server or across multiple indexing machines), Mímir offers *federated indexes*, where several identically-configured indexes are presented to clients as a single virtual index. The indexes that make up the federation are independent, and if documents for indexing are shared out evenly between them then indexing performance should scale up linearly with the number of component indexes, as shown in figure 4.3.

As well as using federated indexes to parallelise the indexing process, it is also possible to build a "longitudinal" federated index. Since indexing speed will inevitably decrease over time as more annotations are added to an index, it may be desirable to restart with a fresh index from time to time. A collection of indexes, one per month, for example, could be presented as a single federated index covering a whole year.

4.4 Conclusions

It is possible to achieve throughput of several million Tweets per hour when indexing Tweets in Mímir, but careful choice of which annotation features to index is necessary to ensure adequate performance. In particular, it is important to avoid indexing features that are likely to be unique or nearly so. Mímir performs best when operating on features with a closed set of possible values, where the same value is shared by many different annotations. Indexing Tweet IDs leads to particularly poor performance, and timestamps should be stored at the coarsest granularity possible for a given use case (e.g. the nearest day or hour rather than the nearest second). If a single index cannot provide sufficient throughput, then several independent indexes can be federated, allowing performance to scale linearly with the number of parallel indexes.

We also performed a separate user-based evaluation experiment, as part of the EnviLOD project, which used the LODIE DBpedia-based entity recognition and linking system from TrendMiner [AGBP13, AGB$^+$12] to annotate semantically grey literature from the domain of environmental science. The resulting documents, annotations, and URIs were indexed in Mímir and it was configured to execute SPARQL queries against two semantic repositories (one for GeoNames and one for DBpedia). The user-based evaluation results are reported in [TBRC].

Overall, the user-based evaluation concluded that there are no major issues with using bespoke form-based user interfaces for semantic search with Mímir. Almost all users learnt to use integrated semantic search successfully after a short demonstration. In more detail, 9 of the 16 (56.3%) participants agreed or strongly agreed that they would use the form-based semantic search UI frequently. Another 6 participants were neutral and only 1 participant strongly disagreed. 14 of the participants (87.5%) disagreed or strongly disagreed with the statement that the semantic search UI is unnecessarily complex and 2 were neutral.

The rationale behind carrying out the user-based evaluation outside of rethe TrendMiner use cases is that, through the British Library, we were able to gain access to a larger number of participants, with diverse skills and areas of work. However, since we used the TrendMiner tools unchanged, the results are representative of other application domains, including the two studied in TrendMiner.

Chapter 5

Conclusion

This deliverable described the evaluation of the summarisation for social media work introduces in Deliverable 4.3.1. We have performed more appropriate automatic evaluation using a wider variety of scoring measures, and have introduced a new manual evaluation of our top performing system and several baselines. The evaluation carried out shows our system to perform worse in manual evaluation than random baselines, though it is difficult to determine the significance of these results due to weaknesses in our experimental set up.

We have begun to investigate the differences in performance, and found that the preferences of SORA have indeed changed for this particular task. Further qualitative investigation would be required to explain these differences, though we assert that to some extent is appears that the summary evalution task for this domain remains extremely subjective, and difficult to evaluate accurately where only a single annotator is available.

5.1 Ongoing work

The disagreement between contemporary judgements by SORA, and SORA one year later, has demonstrated that the task is even more subjective than we initially believed. As such, we would like to continue the work by corroborating more of the historical judgements with annotators from the University of Sheffield. By creating a data set with relevance judgements from many more users, we hope to develop a more robust testing set which will hopefully allow stronger automatic evaluation without overfitting.

We also acknowledge the weaknesses in our manual evaluation of timeline summaries, where five different approaches were shown to users and scored on an ordered catagorical scale. By reducing the number of approaches under test and allowing a simple preference, we should be able to state with greater confidence which approach is strongest in each dimension. We envision a test setup where we simply compare the best approach according to automatic evaluation, augmented for redundancy prevention, against the status quo of ordering tweets reverse chronogically.

Appendix A

Tweet Annotation and Summary Evaluation Guidelines

Here we include the exact wording used in the data collection and user study tasks.

A.1 Gold Standard Tweet Annotation Guidelines

A.1.1 Filter creation

Screenshot in Figure A.1

Please help us to make sense of your timeline by choosing some words or usernames below that are related to each other.

These terms may discuss a related subject, or relate to the same place. They might all be about the same news item or the come from same group of friends. For example, It might be possible to relate “Pancake”, “Valentine’s”, “Batter” and “Flip” as they all relate to events in the same week.

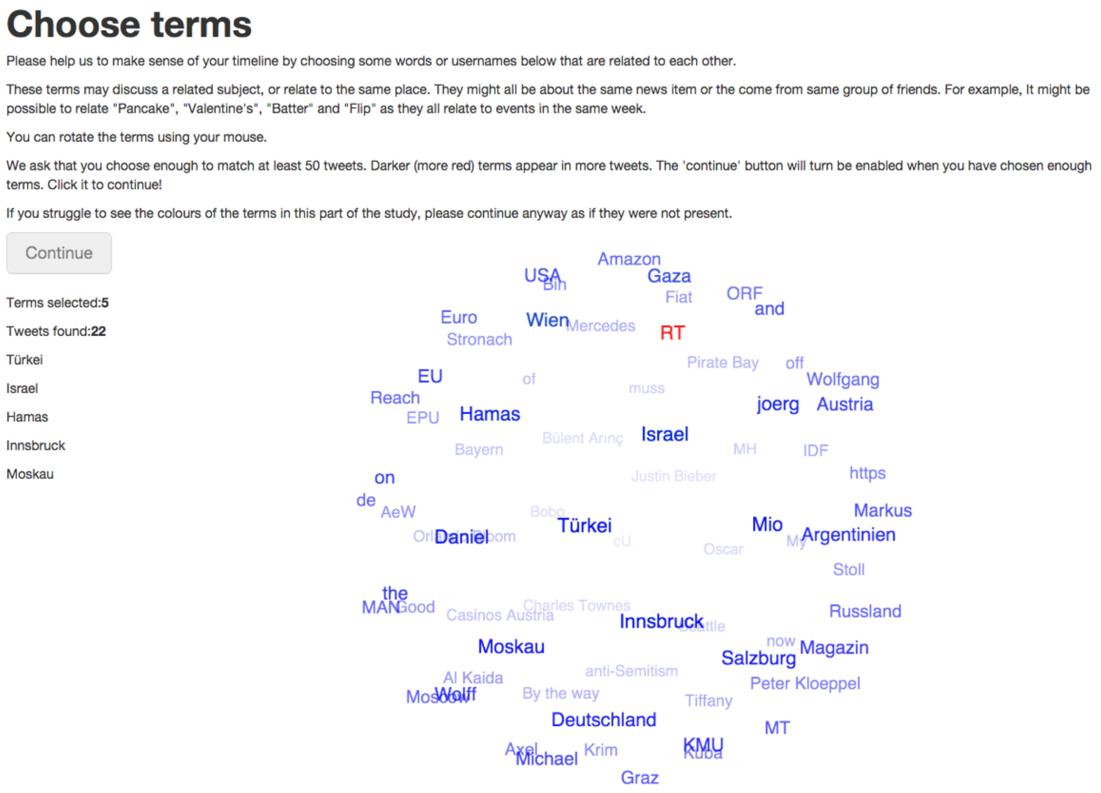
You can rotate the terms using your mouse.

We ask that you choose enough to match at least 50 tweets. Darker (more red) terms appear in more tweets. The ‘continue’ button will turn be enabled when you have chosen enough terms. Click it to continue!

If you struggle to see the colours of the terms in this part of the study, please continue anyway as if they were not present.

A.1.2 Judging Relevance

Screenshot in Figure A.2



Choose terms

Please help us to make sense of your timeline by choosing some words or usernames below that are related to each other.

These terms may discuss a related subject, or relate to the same place. They might all be about the same news item or the come from same group of friends. For example, It might be possible to relate "Pancake", "Valentine's", "Batter" and "Flip" as they all relate to events in the same week.

You can rotate the terms using your mouse.

We ask that you choose enough to match at least 50 tweets. Darker (more red) terms appear in more tweets. The 'continue' button will turn be enabled when you have chosen enough terms. Click it to continue!

If you struggle to see the colours of the terms in this part of the study, please continue anyway as if they were not present.

Continue

Terms selected:**5**

Tweets found:**22**

Türkei

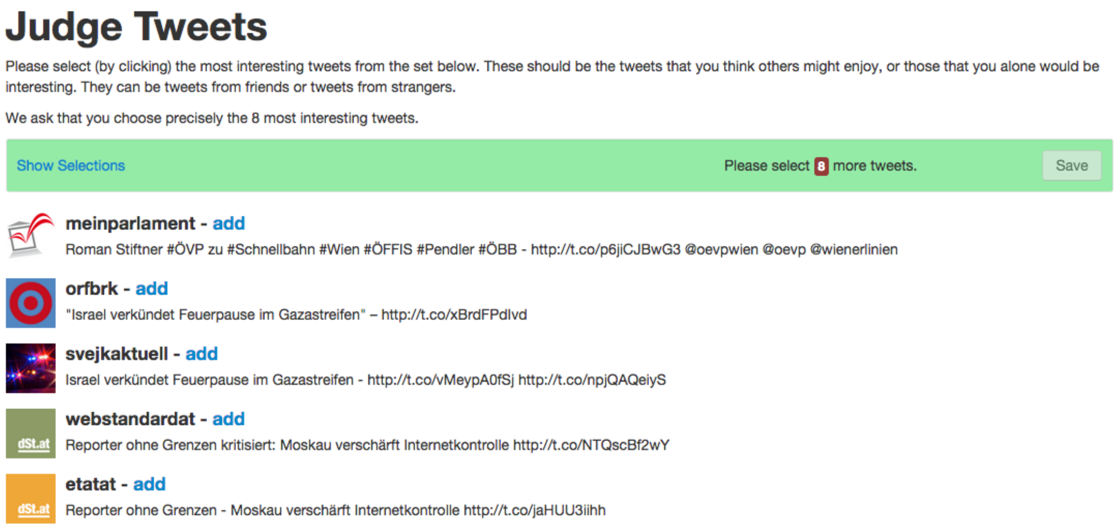
Israel

Hamas

Innsbruck

Moskau

Figure A.1: Creating a tweet filter



Judge Tweets

Please select (by clicking) the most interesting tweets from the set below. These should be the tweets that you think others might enjoy, or those that you alone would be interesting. They can be tweets from friends or tweets from strangers.

We ask that you choose precisely the 8 most interesting tweets.

Show Selections | Please select 8 more tweets. | Save

meinparlament - add
Roman Stiftner #ÖVP zu #Schnellbahn #Wien #ÖFFIS #Pendler #ÖBB - http://t.co/p6jiCJBwG3 @oevpwien @oevp @wienerlinien

orfbrk - add
"Israel verkündet Feuerpause im Gazastreifen" – http://t.co/xBrdFPdlvd

svejkaktuell - add
Israel verkündet Feuerpause im Gazastreifen - http://t.co/vMeypA0fSj http://t.co/npjQAQeiyS

webstandardat - add
Reporter ohne Grenzen kritisiert: Moskau verschärft Internetkontrolle http://t.co/NTQscBf2wY

etatat - add
Reporter ohne Grenzen - Moskau verschärft Internetkontrolle http://t.co/jaHUU3iihh

Figure A.2: Original tweet relevance annotation interface

Please select (by clicking) the most interesting tweets from the set below. These should be the tweets that you think others might enjoy, or those that you alone would find interesting. They can be tweets from friends or tweets from strangers.

We ask that you choose precisely the 8 most interesting tweets.

A.2 Manual Summary Evaluation Guidelines

Screenshot in Figure 3.1

Please read first the set of original tweets. As you continue to scroll your browser window, you will see summary A, followed by several brief questions for you to respond to. Aterwards, please scroll and consider summaries B, C, D, and E.

Please give your opinions on the following regarding the summary above:

- The summary captures all the important information from the full set of tweets
- There were several tweets in the summary that were repeating very similar information
- I could use this summary to study political figures or events
- Given your responses above, please rate this summary as a whole

A.3 Reannotating Tweets for Interestingness

Screenshot in Figure 3.2

Please select (by clicking) the most interesting tweets from the set below. We ask that you choose precisely the 8 most interesting tweets.

Bibliography

[AGB$^+$12] Niraj Aswani, Mark Greenwood, Kalina Bontcheva, Leon Derczynski, Julian Moreno Schneider, Hans-Ulrich Krieger, and Thierry Declerck. Multilingual, ontology-based information extraction from stream media - v1. Technical Report D2.2.1, TrendMiner Project Deliverable, 2012.

[AGBP13] Niraj Aswani, Genevieve Gorrell, Kalina Bontcheva, and Johann Petrak. Multilingual, ontology-based information extraction from stream media - v2. Technical Report D2.2.2, TrendMiner Project Deliverable, 2013.

[BNJ03] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, March 2003.

[CG98] Jaime G. Carbonell and Jade Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Research and Development in Information Retrieval*, pages 335–336, 1998.

[DDF$^+$90] S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41:391–407, 1990.

[IK11] David Inouye and Jugal K. Kalita. Comparing Twitter summarization algorithms for multiple post summaries. In *SocialCom/PASSAT*, pages 298–306, 2011.

[Lik32] Rensis Likert. A technique for the measurement of attitudes. *Archives of psychology*, 1932.

[Lin04] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In Stan Szpakowicz Marie-Francine Moens, editor, *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[MRS08] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval.* Cambridge University Press, New York, NY, 2008.

[MT04] R. Mihalcea and P. Tarau. TextRank: Bringing order into text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 404–411, 2004.

[NP04] Ani Nenkova and Rebecca Passonneau. Evaluating content selection in summarization: The pyramid method. In *Proceedings of HLT-NAACL*, pages 145–152, 2004.

[Por01] Martin Porter. Snowball: A language for stemming algorithms, 2001.

[RJST04] Dragomir R. Radev, Hongyan Jing, Malgorzata Stys̆, and Daniel Tam. Centroid-based summarization of multiple documents. *Information Processing and Management*, 40(6):919–938, November 2004.

[SG11] Guy Shani and Asela Gunawardana. Evaluating recommendation systems. In *Recommender systems handbook*, pages 257–297. Springer, 2011.

[SHK10] B. Sharifi, M. A. Hutton, and J. Kalita. Summarizing Microblogs Automatically. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 685–688, Los Angeles, California, June 2010.

[TBRC] V. Tablan, K. Bontcheva, I. Roberts, and H. Cunningham. Mímir: an open-source semantic search framework for interactive information seeking and discovery. *Journal of Web Semantics*.

[UC11] Ibrahim Uysal and W. Bruce Croft. User oriented tweet ranking: a filtering approach to microblogs. In *Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011, Glasgow, United Kingdom, October 24-28, 2011*, pages 2261–2264, 2011.

[YLL12] Rui Yan, Mirella Lapata, and Xiaoming Li. Tweet recommendation with graph co-ranking. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 516–525, Jeju Island, Korea, 2012.