Examining New Event Detection

Johannes Schanda
Information School
The University of Sheffield
S1 4DP, Sheffield, UK
joschanda@schanda.de

Mark Sanderson
School of CS & IT
RMIT University
Melbourne, Australia
mark.sanderson@rmit.edu.au

Paul Clough
Information School
The University of Sheffield
S1 4DP, Sheffield, UK
p.d.clough@shef.ac.uk

ABSTRACT

We examine the accuracy of first story detection on traditional news collections and on a re-purposed source of academic material. The impact on accuracy of detecting an early rather than the first story is examined, showing that accuracy increases under a broader time window, however, the increases on some collections are small. Even on collections where the increase is large, many new events are still missed and there remains an underlying challenge to detecting new events. An analysis of temporal and vocabulary profiles of topics within their source collections is conducted. Analysis of the results establish the underlying causes of the patterns seen in the experimental results with respect to the different source types and performance. The usefulness of new criteria for new event detection and success across source types is discussed.

1. INTRODUCTION

Spurred by widespread interest in organizing and filtering information more effectively, a coordinated attempt was made to develop Topic Detection and Tracking (TDT) systems. In 1998, Allan et al. [5] spawned a field of research centered on the annual TDT workshop, a subtask of which was First Story (or later New Event) Detection. This task required systems to identify, the first *story* published on a new *topic*. The difficulty of having only one chance for a correct decision of finding the first story resulted in relatively poor accuracy for systems on New Event Detection. The reasons for this were examined in detail by Allan et al. [8]. By the end of the TDT workshops in 2004, New Event Detection accuracy was not significantly improved.

While detection of events is actively researched in domains such as email or social media [2], there are still parts of the original TDT work that have not been fully explored, which are the subject of this paper. First, we examine the level of increase in detection accuracy when the requirement to find the very first story is relaxed. Such an analysis would provide us with an understanding of the extent to which relaxing the evaluation criteria has the potential to improve accuracy, and at what point the quality of the classifier becomes the limiting factor as stipulated in [8]. Second, we extend the TDT methodology to a new domain, academic publications. Here, new lines of research (new events) occur, however the requirement to locate the first publication is not as important as it was in the news domain that TDT focused on. The research questions of this paper are, therefore:

- How does new event detection accuracy increase when the time window in which a new event can be found is widened?
- Does the relationship between new event detection accuracy and time window size change when TDT is applied in a different domain?
- What cause any differences between the domains?

Evidence is presented of the improvements in accuracy which are gained by widening the time window in which a detected story may be considered a successful detection of a new event. These experiments are conducted on both traditional TDT news collections and an academic paper collection. We establish a foundation on which decisions about what windows to use in new event detection systems can be made and to quantify the improvement attainable by widening the time windows. We then examine the temporal and vocabulary profiles of topics within their respective collections to quantify the differences between news and academic source text.

The paper introduces previous work (Section 2), followed by Section 3, which discusses the experiment design. Sections 4 and 5 present experimental results and analysis, with discussion in section 6. Section 7 concludes the paper.

2. PREVIOUS WORK

The evaluation of TDT research was formalized by Allan et al. [5] who defined the measures *miss rate* and *false alarm rate*. Miss rate is the ratio of stories that should have been detected as indicators of a new event (but were not) to the total number of stories that herald a new event. False alarm rate is the ratio of stories incorrectly detected as new events to the total number of stories detected as a new event. Being error rates, it is desirable to minimize both.

Allan et al. [6] explored different similarity measures, as well as clustering methods, to determine if a story is part of an existing event or describing a new event. Agglomerative clustering was explored, as well as a k-nearest neighbour

ADCS '14 November 27-28 2014, Melbourne, VIC, Australia

http://dx.doi.org/10.1145/2600428.2609469.

approach. Stokes et al. [22] studied the value of representing documents as sequences of lexical chains as opposed to terms, but with limited success. Brants et al. [11] experimented with generating TF-IDF models specific to particular events, based on the assumption that different events employed substantially different vocabularies and so the differentiating power of the model would be enhanced by using an event-specific vocabulary. The vector space model was extended to utilize named entities [7]. Another common approach was to utilize a variety of probabilistic models to estimate the distance between a story and a given event [16]. Leek et. al. [17] explored such measures, experimenting with a Hidden Markov Model (HMM) to simulate the contributions different topics made to a given story. Zhang et. al. [24] investigated using a tree index of stories in order to speed up the process of assigning a story to a topic.

While error rates were reduced over time, new event detection remained hard. Allan et. al. examined this problem in some detail [8]. They found that TDT techniques treated new event detection as a special case of tracking, which rendered them vulnerable to compounding errors over time, as each incorrect assignment of a story resulted in distortion of the topic definitions. This made it increasingly difficult to assign stories correctly. The paper argued that in order to improve new event detection accuracy, tracking performance (i.e. quality of the classifiers) would have to be improved by a factor of approximately twenty.

There have been a number of attempts to go beyond the traditional definition of new event detection. These can be split into two trends. First, aggregating detection across multiple source types in a coherent fashion (e.g. [3]), where the goal is to build story lines which cross media boundaries. This also allows examining the temporal profiles of stories as they develop. The results reported are comparable to those of other TDT systems on the new event detection task. Second, applying TDT principles to new domains, or the addition of specific new criteria to the task. Domains considered are varied, and include security [4], Tweets [18], and videos [21, 12]. An example of this trend is presented in a paper [4] which aims to extract event signatures in order to track terrorism news. The results reported in this paper do not use the standard TDT evaluation measures, which makes it difficult to compare their results to other TDT systems.

3. EXPERIMENTS

The experiments were designed to quantify the reduction in error rates that can be achieved when new event detection is redefined as detection within a given time window. We start by describing the collections.

3.1 Collections

The first corpus used is the TDT4 corpus [23], chosen for its extensive use in TDT evaluation. The corpus comprises a variety of types of news text. It includes both newswire stories and transcribed spoken text from a variety of sources. It is segmented into three different test collections: TDT_spoken a collection of English transcribed audio; TDT_written a collection of English written news; and TDT_eng the union of the first two collections. Table 1 lists the segmented collections, including identifiers for each source contained in the collection.

The purpose of segmenting the TDT4 corpus into two subsets was to enable the examination of the effects of different types of source material on the performance of the detection system. Material that originates from a written form is different in vocabulary and sentence length to that transcribed from speech[5]. Table 2 shows statistics for the TDT4 collections.

3.2 Academic Sources

New Event Detection is not only conducted on news. Another example of tracking behavior can be found in academia where researchers scan for new articles on a chosen topic. We decided therefore, to find a collection that could be adapted to be used as an academic TDT collection. The requirements of such a collection are to have a set of documents, some of which are marked up as describing a particular event. The only other requirement of a TDT collection is that the documents are time stamped so that they can be processed to simulate a stream of data. The qrels of a retrieval test collection could be said to be similar to the marked up documents in a TDT collection: they list which documents are relevant to a topic. While they are not the same, we judged that topics and events were sufficiently similar to allow retrieval test collections to act as proxies to TDT collections.

The TREC Genomics retrieval collection [1] consists of abstracts of medical papers, each with a date of publication. Additionally, topics in the Genomics corpus cover a range of medical and genetic subjects and can be treated as events for a TDT experiment. Details of the collection *gen_small* used for these experiments are presented in Table 3.

3.3 Time Windows

The primary experimental variable was a differing time window. It was found that the size of the time windows had to be different for the two collections. For TDT, the windows of 0M (zero minutes, i.e. first story detection), 30M, 1H, 6H, 12H, 1D, 2D, 7D, 14D and 30D were used. For the Genomics collection, 0M, 1D, 7D, 14D, 30D, 60D, 120D and 360D reflected the greater lifetime of topics.

3.4 Evaluation Criteria

The use of time windows required adapting the existing evaluation criteria. As the TDT system under evaluation is run, a series of story clusters are formed around possible events e. Evaluation is carried out by examining each story in each cluster in publication timestamp order. If a story is relevant to a particular event and it is published within the stated time window x of the first story, and there is not already a cluster that begins with a story deemed as relevant to this event, then the story is judged a successful detection of a new event. If there is already a cluster which begins with a story determined to be a successful detection, then this story is a false alarm as it should be assigned to that cluster. If the story is not published within the evaluation time window, it is also a new event false alarm. Any event which does not have a cluster beginning with a story deemed a success after all clusters have been examined is deemed to be a miss.

3.5 System and Method

A new event detection system was constructed based on the Terrier IR system [19]. For each new story to be classified, a document feature vector was created. Stopwords were removed, and Porter stemming applied [20]. Term weighting was calculated using TF-IDF and document comparison

Identifier	Source Tags
TDT_spoken	CNN_HDL, ABC_WNT, NBC_NNW, VOA_ENG, MNB_NBW
TDT_written	APW_ENG, NYT_NYT
TDT_eng	APW_ENG, NYT_NYT, CNN_HDL, ABC_WNT, NBC_NNW, VOA_ENG, MNB_NBW

Table 1: Collections derived from TDT4

Statistic	*TDT_eng*	*TDT_spoken*	*TDT_written*
Stories	25,502	10,412	15,090
Annotated Stories	1,181	498	683
Annotated Topics	37	34	33
Vocabulary	78,290	26,039	71,216
Size	74 MB	12 MB	62 MB

Table 2: Statistics for TDT collections

used the cosine similarity measure.

The system uses a first-k clustering algorithm, which is a variation of nearest neighbor clustering with $k = 1$, where a sliding window is applied to the candidate documents. The distance of the nearest story is then used to make a classification decision for the new story. If it is closer than a threshold, the story is assigned to the same event. Otherwise, a new event is generated and the story is assigned to it as the first story. Working through the distance threshold space generates results at various levels of similarity for New Event Detection.

3.6 Presentation of Results

The results will be shown graphically using DET (Decision Error Trade) curves [6]: a visual representation of the tradeoff in false alarm and miss rates throughout the detection confidence space. Detection confidence is the variable which determines how "sure" the system must be that a given story describes a new event in order to class it as such.

Normalized detection cost functions are a method of assessing a system's performance as a single number [6]. Two functions will be used. The traditional method denoted by $C_{DetNormAvg}$ [6] assigns a cost to both a missed detection and a false alarm. These costs are set depending on the TDT task attempted. The actual values assigned to these costs are less important than their consistent use. Our values are taken from [14].

We also use a new metric, denoted by $CumErr_{Min}$, to represents the minimum cumulative error generated by the system using a particular configuration. It is the minimum of adding the error rates at each threshold. It is defined as follows:

$$CumErr_{Min} = \min(\{CumErr_I | I \in ThresholdSpace\})$$

where

$$CumErr_I = P_{MissI} + P_{FaI}$$

$I :=$ Threshold value

$P_{MissI} :=$ Probability of missed detection at threshold I

$P_{FaI} :=$ Probability of false alarm at threshold I

The measure gives an indication of system performance. P_{Miss} and P_{Fa} vary individually throughout the threshold space. This measure finds the point at which the sum of errors is at its lowest level in each threshold space. The difference variance of the $CumErr$ results for each time window are analyzed for significance using the F-Test [10].

Statistic	Value
Stories	500,000
Annotated Stories	647
Annotated Topics	47
Vocabulary	132,431
Size	3,335 MB

Table 3: Statistics for gen_small collection

Time window	$C_{DetNormAvg}$ TDT-_eng	TDT-_spoken	TDT-_written	$CumErr_{Min}$ TDT-_eng	TDT-_spoken	TDT-_written
0M	0.47	0.47	0.38	0.29	0.31	0.31
30M	0.47	0.47	0.38	0.23	0.28	0.28
1H	0.47	0.47	0.38	0.23	0.28	0.28
6H	0.47	0.47	0.38	0.23	0.28	0.28
12H	0.47	0.47	0.38	0.23	0.28	0.28
1D	0.47	0.47	0.38	0.23	0.28	0.28
2D	0.47	0.47	0.38	0.23	0.28	0.28
7D	0.47	0.47	0.38	0.23	0.28	0.28
14D	0.47	0.47	0.38	0.23	0.25	0.24
30D	0.47	0.47	0.38	0.20	0.22	0.22

Table 4: Evaluation for the TDT4 collections. Compared to 0M no differences are significant.

4. RESULTS

Table 4 shows the $C_{DetNormAvg}$ and $CumErr_{Min}$ values for all three TDT collections; there are two primary effects in these results. The differences in accuracy across the time windows, and the differences across collections. The widening of time windows barely caused an increase in accuracy. Large windows such as 30 days (30D) appear to have some impact, though none of these improvements, compared to 0M, were statistically significant.

The effect returned by the addition of time windows may be rooted in the characteristics of the classifier. If an algorithm misses a topic when it is first published, it will miss it again at a similar threshold when it is published later by a different source. Even if the material originates from a different source, the vocabulary that marks a story as dealing with a new event may not change substantially, hence the algorithm will keep missing it. This effect is discussed further in Section 5.

Smaller time windows have little effect: $C_{DetNormAvg}$ and $CumErr_{Min}$ for all windows between 30 minutes and 1 day are the same across the collections. This suggests there is some benefit to allowing a small window for missing detection, such as 30 minutes. However, there is no benefit in further relaxing the window to allow for larger amounts of time, such as 7 days.

Time window	$C_{DetNormAvg}$	$CumErr_{Min}$	p value of an F-Test v. 0M on $CumErr$
0M	0.58	0.64	
1D	0.58	0.58	0.34
7D	0.58	0.55	0.26
14D	0.58	0.50	0.03
30D	0.57	0.45	0.00
60D	0.57	0.38	0.00
120D	0.57	0.36	0.00
360D	0.57	0.33	0.00

Table 5: Evaluation for gen_small.

It was found that for some events in the collection, the gap between the first and second story was long, which may indicate why a small widening of the time window did not have a large effect. It may be that the effect of the results being identical between the 30 minute and 1 day window is a product of how the topics are annotated in this corpus. This will be explored in Section 5, along with the different ways various topics react to the addition of time windows.

4.1 Academic Sources

Examining the TREC Genomics collection, Table 5 shows the $C_{DetNormAvg}$ and $CumErr_{Min}$ values as well as the F-Test results across the time windows. These results are markedly different from those of the TDT corpus based collections. The absolute performance, i.e. the absolute error rates, on this collection are somewhat worse than on the TDT4 based collections. This can be explained by the greater size of this collection, which results in a smaller proportion of stories being relevant and thus increasing the odds of making mistakes (see Tables 3 and 2). The proportion of annotated stories in this collection is 0.001% whereas on *TDT_written*, it is 0.045%.

The main observation from these results, however, is the improved relative performance from adding larger time windows. The smallest window beyond the traditional first only measure used here, one day, shows an increase of 0.06 on the $CumErr_{Min}$ measure. On the next window the same measure gains an additional 0.03.

The difference between the *0M* and *360D* time windows on the $CumErr_{Min}$ measure is 0.31, which is roughly half of the *0M* measure of 0.64. This is partially due to the pattern of publication in the TREC Genomics corpus. Since many stories are collected from academic journals and proceedings, allowing a one day window opens all the stories published in the same issue to be considered. Section 5 examines the publication patterns in various topics.

Table 6 shows a comparison of evaluation functions for the various collections for the traditional measure (*0M*) and various time windows. These results show the differing characteristics of the two collections. At the *0M* time window, the *gen_small* collection performs significantly worse than any of the TDT corpus based collections. At *30D* this difference is already less pronounced, and at *360D*, the largest time window allowed for on *gen_small*, compared to *30M* as the largest evaluated window for the TDT based collections, the difference is smaller still.

Time window	Collection	$C_{DetNormAvg}$	$CumErr_{Min}$
0M	TDT_eng	0.47	0.29
	TDT_spoken	0.47	0.31
	TDT_written	0.38	0.31
	gen_small	0.58	0.64
30D	TDT_eng	0.47	0.20
	TDT_spoken	0.47	0.22
	TDT_written	0.38	0.22
	gen_small	0.57	0.45
360D	gen_small	0.57	0.33

Table 6: Comparing TDT4 and genomics collections

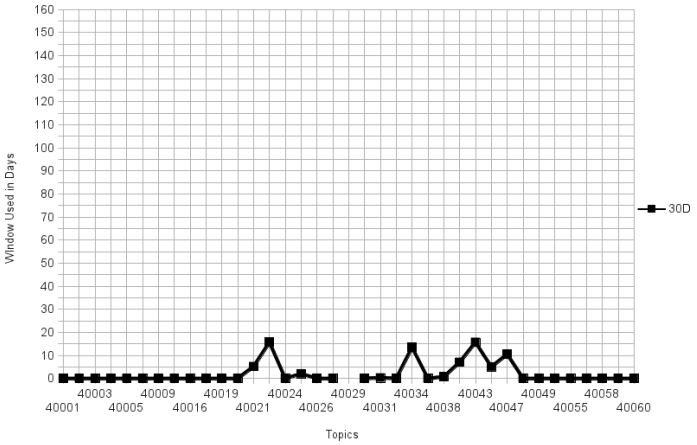

Figure 1: Best time window found for each topic in TDT_eng. Missing data are topics with only one relevant document or which have not been detected at the maximum window.

5. TOPIC PROFILES

Given the variability of impact of the time windowing approach established above, an examination of the nature of the topics in the two collections was undertaken to attempt to understand why one was more helped by windowing than the other. The *maximum topic time window* was defined as the time window for which the lowest $CumErr_{Min}$ was measured. The feature shows on a topic by topic basis which time window was the best to use, see Figures 1 and 2. As can be seen for *TDT_eng* the impact of time windows was small and simply detecting the first story was the most effective approach to minimizing $CumErr_{Min}$. In Figure 2, one can see that there are many more topics for which the $CumErr_{Min}$ is reduced; on the *gen_small* collection there is more impact of time windows across the topics.

In order to better understand the reason for the differences between the collections, the *temporal profile* (the pattern with which associated stories are published over time) and the *vocabulary profile* (the characteristics of the words associated with a topic within its collection) were examined. A series of features and measures were explored, which are detailed in the following sub-sections.

5.1 The 1^{st} and 2^{nd} Story Gap

Figures 3 and 4 show the gap, in days, between the first

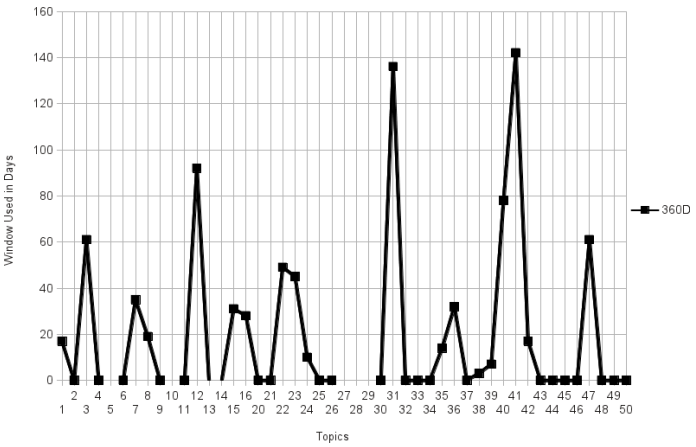

Figure 2: Best time window found for each topic in gen_small. Missing data are topics with only one relevant document or which have not been detected at the maximum window.

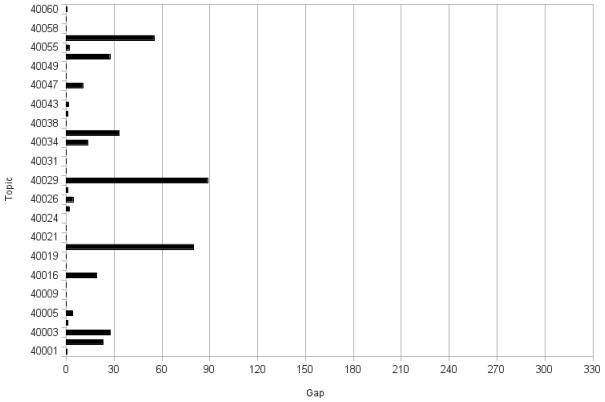

Figure 3: Time gap (in days) between 1^{st} and 2^{nd} stories in TDT_eng. Topics with no bar have only one associated story or the gap is zero days.

and second story for the topics in *TDT_eng* and *gen_small* respectively. This feature determines what time window will be required at a minimum to enable detection of a later story as part of a new event. Many of the topics in *gen_small* have gaps measured in multiple days, this contributes to many of the wider time frames not being more helpful. For the topics in *TDT_eng*, the majority have a gap of less then one day, though a notable minority of topics have, perhaps unexpectedly, much larger gaps.

5.2 First 25% of Stories

This feature was examined to characterize the 'burstiness' or 'eveness' of a topic's stories over time. Across the period of time that stories are published on a particular topic, the proportion of time that passes before 25% of the stories of that topic were published was determined. Figures 5 and 6 present histograms for *TDT_eng* and *gen_small* respectively.

The histograms for the two collections are quite different. The *TDT_eng* collection has many more topics in the 5% bin than *gen_small*, tailing off quickly. The *gen_small* collection

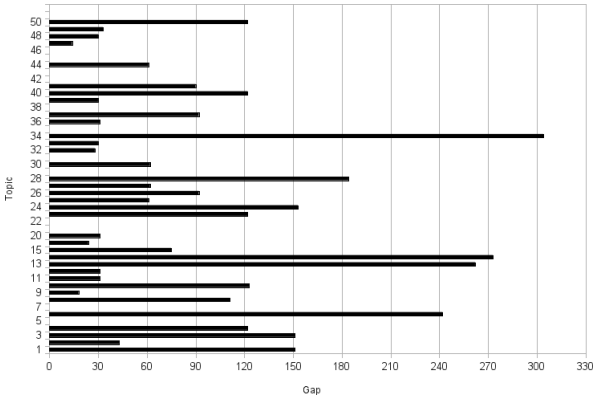

Figure 4: Time gap (in days) between 1^{st} and 2^{nd} stories in gen_small. Topics with no bar have only one associated story or the gap is zero days.

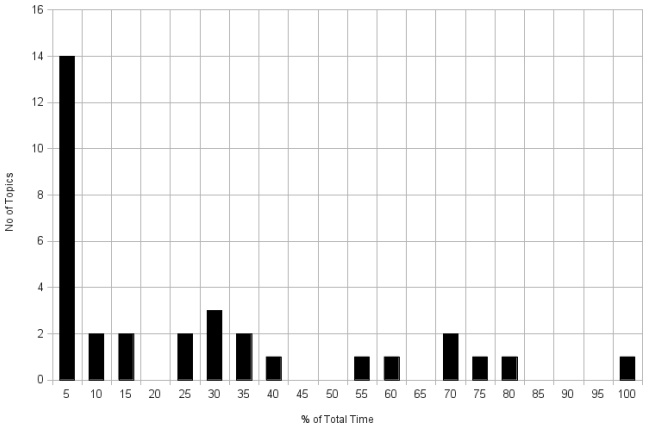

Figure 5: The % time till 25% of stories were published - for *TDT_eng*. Note only topics with > 1 story are plotted here.

shows a more even distribution with a peak at 40%. The differences provide an indication of why the widening of the time window reduced the error on the Genomics collection more than on the TDT based collections.

5.3 Story/Time Distributions

Another way of presenting the time profile of topics is to examine the relative story/time distribution for each topic, i.e. the proportion of stories associated with a topic that are published over the total time associated with a topic. Figures 7 and 8 show the proportion of stories published within each 10% of a topic's lifetime. The data is presented as a stacking chart. One can see in Figure 8 that on the *gen_small* collection stories are much more evenly distributed than on the *TDT_eng* collection. Again, one can see why the impact of the wider time windows continues to grow on the *gen_small* collection: there are many more topics with a relatively even coverage of stories over the lifetime of the topic.

5.4 Vocabulary Features and Window Performance

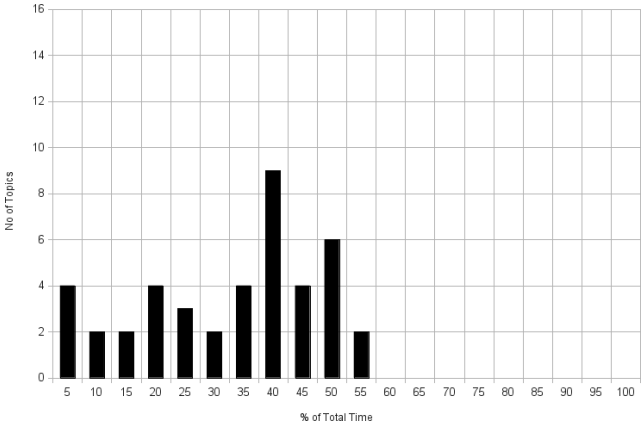

Figure 6: The % time till 25% of stories were published - for *gen_small*. Note only topics with > 1 story are plotted here.

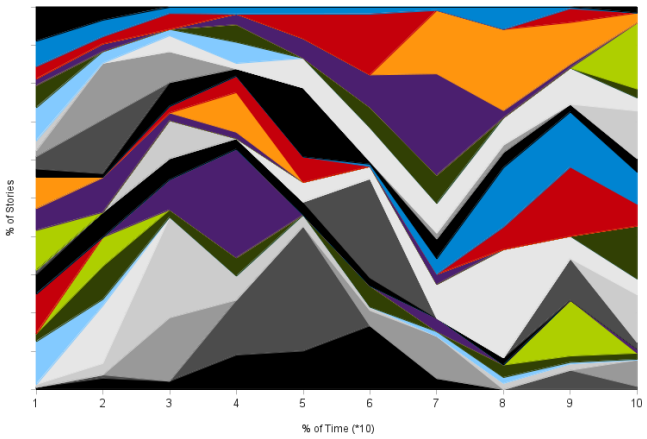

Figure 7: Volume of per topic story publication over time in the *TDT_eng* collection.

Topics are differentiated from one another by their vocabulary. One of the ways in which this can be measured is by comparing the language models describing the topics with that describing the collection in general. The language models were compared using the measures of perplexity [15] and entropy, as generated with the CMU Statistical Language Modeling Toolkit [13].

Perplexity and entropy are measures from information theory. In the case of information retrieval, entropy usually measures the uncertainty of a language model with respect to the text it describes. Low entropy of a language model indicates that is has a low degree of uncertainty; the model describes its text well. Perplexity quantifies the degree of confusion a model has to a particular test sample. Better models assign a higher probability to the test samples, which results in lower perplexity.

Here these measures are used to quantify the uncertainty of a Unigram language model of the vocabulary of the entire collection with respect to the text of all the stories associated with a particular topic. It was found that the perplexity of topics is much higher on the *gen_small* collection than on

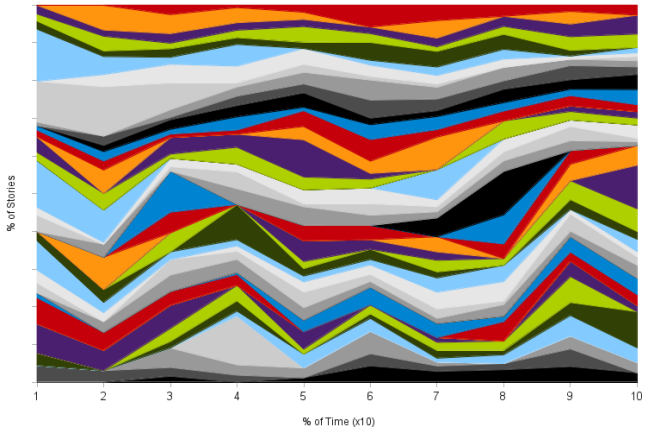

Figure 8: Volume of per topic story publication over time in the *gen_small* collection.

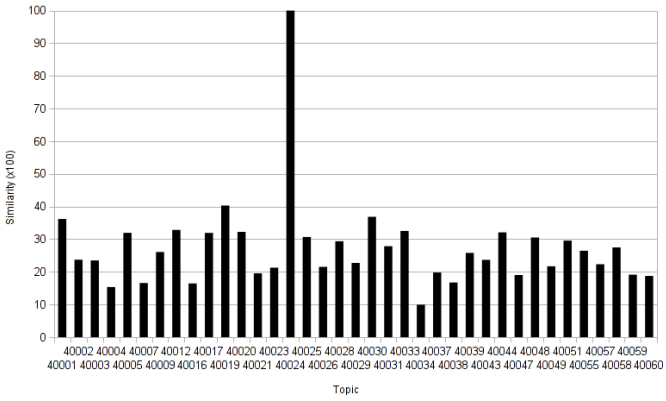

Figure 9: Pairwise Similarity in TDT_eng topics

TDT_eng, indicating that the topic text diverges more from the general collection than it does on *TDT_eng*. This may seem surprising given the previous results showing that it is significantly more difficult to detect topics on *gen_small*. Again, this difficulty can be attributed to the difference in size of the two collections.

The mean pairwise similarity between stories associated with each topic is another way to measure the vocabulary differences between topics and collections. For the most part the pairwise similarity in topics is relatively similar across each collection, as shown in Figures 9 and 10. Much like for perplexity and entropy, similarity is higher across all topics on the *TDT_eng* collection. There are a three outlier topics with a similarity of 100, which are topics with only one associated story; they are completely similar within themselves.

5.5 Correlating Features with Accuracy

The correlation between a set of features described in the section before and the impact of time windowing was computed, see Table 7. It can be seen that there is a substantial difference in behavior between the Genomics and TDT4 based collections. On the *TDT_eng* collection, none of the correlations are significant. In contrast, the *gen_small* collection has four features which correlate relatively strongly

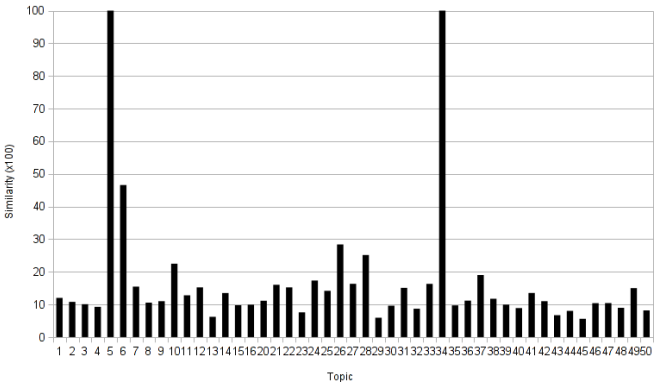

Figure 10: Pairwise Similarity in gen_small topics

with the impact that time windows have on the accuracy of New Event Detection. These features are the number of stories, % of time to 25% of stories, topic lifetime, and perplexity.

A correlation was also found for topic lifetime and the improvement in accuracy of new event detection as a time window grows. This was to be expected. The correlation of the number of stories in a topic is most likely due to this number being linked to the lifetime of a topic. The examination of the vocabulary profile using perplexity shows that the distinctiveness of the stories of a topic correlates with the impact of a time window. This is perhaps a little surprising as although one might expect vocabulary to impact on overall accuracy of new event detection; seeing that it impacts on the value of time windows was unexpected.

6. DISCUSSION

The addition of time windows to event detection reduces error rates, though only significantly for the *gen_small* collection. Nevertheless, for both corpora, the remaining error is still high (0.20 for *TDT_eng* and 0.33 for *gen_small*).

All topics (except one) that are detected at any point in the range of time windows are detected at some point in the threshold space, even on the traditional first story only measure. What changes across the time windows is which topics are detected at each point in the threshold space. If a topic, which is detected at a lower point in the threshold space on the traditional measure, has an associated story within a time window which can be detected at the current threshold point, that topic's performance will increase as that time window is reached. Working through the threshold space trades off successful detection of harder to find stories with increasing the false alarm rate, where stories start topics which should not. By working through the threshold space we can find the point of minimum error, where the sum of both miss and false alarm rates is the lowest. Thus the overall accuracy of the system is improved if more topics are detected at the point of minimum cumulative error.

Since essentially all the topics which are detected at any point in the time window are detected at the traditional measure and stories associated with a topic do not tend to become much easier or harder to detect over time the performance of the classifier is still the limiting factor for the systems performance. Essentially, if the actual first story is missed, subsequent ones are likely to be missed also.

Section 5.5 examined feature accuracy correlations. There is a stronger correlation with the vocabulary based feature (e.g. perplexity) on the Genomics collection over the TDT collection. This shows that the topic vocabulary diverges more from the general vocabulary on Genomics. This should make stories easier to detect. However, the absolute accuracy on the Genomics collection is worse than on TDT based collection because of its much greater size. This makes it more difficult to detect the correct stories, but the greater divergence of topic stories from the general vocabulary makes it more likely that there is a subsequent story associated with a topic that can be detected with the introduction of time windows. This explains the greater effect that time windows have on stories *gen_small* collection.

The maximum time windows of 30D and 360D applied to the two collections can be viewed as a best case scenario. They are unrealistically long for a real world application, but they indicate what is possible with the current approach and the current classifiers. The absolute performance in these idealised situations is still not ideal, though more promising on the *gen_small* corpus than on TDT. In the case of detecting new events through long time windows are a realistic case, as many publish monthly or quarterly; therefore the long time windows examined on the Genomics corpus show results interesting to realistic scenarios.

7. CONCLUSION

This paper presented an analysis of the impact of widening the window of time used to detect a new event in a stream of time sorted information. Two different collections were studied in the research: a TDT collection and also a collection of academic articles from the medical domain. The following research questions were posed.

- How does new event detection accuracy increase when the time window in which a new event can be found is widened?
- Does the relationship between new event detection accuracy and time window size change when TDT is applied in a different domain?
- What cause any differences between the domains?

Small increases in the size of time windows are not useful for improving detection accuracy on news collections. However, larger windows improve detection rates for collections of academic papers. Although improvements in accuracy were notable, they were not substantial.

A study of the factors that impact on the accuracy of new event detection was conducted. Here, quite different temporal profiles of story publication were found in the two collections and it was shown that these profiles correlated with detection accuracy. Also the distinctiveness of the vocabulary profile of topic stories was found to correlate with detection accuracy.

As with the impact of time windows, the strength of correlations was quite different on the two collections, emphasizing the importance of examining new event detection in a range of contexts.

8. ACKNOWLEDGMENTS

This work was supported by the Australian Research Council (DP130104007) and by the EPSRC and DSTL.

Feature	Correlation	
	gen_small	TDT_eng
Time gap between first and second story	−0.25*	−0.07
Number of stories in a topic	0.55**	−0.07
The % of time till 25% of stories are published	0.50**	−0.10
Number of stories in the 1st 10% of a topic's lifetime	0.26*	−0.02
The % of stories in the 1st 10% of a topic's lifetime	0.23	0.02
Topic lifetime	0.42**	0.16
Perplexity measured between the vocabularies of the entire collection compared to the vocabulary in the set of topic stories [13]	0.42**	0.23
Mean pairwise similarity between all possible story pairings of the topic [15]	−0.12	0.11

Table 7: Correlation of features to the width of the time window at which topics are detected at minimum error. Significance (two-tailed test using Fisher transformation) at $p < 0.05$ (*), at $p < 0.01$ (**).

9. REFERENCES

[1] TREC Genomics Track. available at `http://ir.ohsu.edu/genomics/`.

[2] C. Aggarwal and K. Subbian. Event Detection in Social Streams. In *SIAM International Conference on Data Mining* , pages 624–635, 2012.

[3] A. Ahmed, Q. Ho, J. Eisenstein, E. Xing, A. Smola, and C. Teo. Unified analysis of streaming news. In *WWW '11*, pages 267–276. ACM, 2011.

[4] S. Ahmed, R. Bhindwale, and H. Davulcu. Tracking terrorism news threads by extracting event signatures. In *ISI '09*, pages 182–184. IEEE, 2009.

[5] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang. Topic detection and tracking pilot study: final report. In *DARPA Broadcast News Transcription and Understanding Workshop*, pages 194–218, 1998.

[6] J. Allan, H. Jin, M. Rajman, C. Wayne, D. Gildea, V. Lavrenko, R. Hoberman, and D. Caputo, editors. *Topic-based novelty detection 1999 summer workshop at CLSP final report*, 1999. available at `http://www.clsp.jhu.edu/ws99/final/Topic-based.pdf`.

[7] J. Allan and G. Kumaran. Text classification and named entities for new event detection. In *SIGIR '04*, pages 297–304, 2004.

[8] J. Allan, V. Lavrenko, and H. Hin. First story detection in TDT is hard. In *CIKM '00*, pages 374–318, 2000.

[9] J. Allan, V. Lavrenko, and R.Swan. Explorations within topic tracking and detection. In J. Allan, editor, *Topic Detection and Tracking; Event-based Information Organization*, pages 197–224. Kluwer Academic Publishers, 2002.

[10] S. Brandt. Statistical and computational methods in data analysis. *Amsterdam: North-Holland, 1976, 2nd rev. ed., 5th repr. 1989*, 1, 1989.

[11] T. Brants and F. Chen. A System for new event detection. In *SIGIR '03*, pages 330–337, 2003.

[12] J. Cao, C. Ngo, Y. Zhang, D. Zhang, and L. Ma. Trajectory-based visualization of web video topics. In *MULTIMEDIA '10*, pages 1639–1642. ACM, 2010.

[13] P. Clarkson and R. Rosenfeld. Statistical language modeling using the CMU-Cambridge toolkit. In *Fifth European Conference on Speech Communication and Technology*, pages 2707–2710, 1997.

[14] J. Fiscus and G. Doddington. Topic detection and tracking overview. In J. Allan, editor, *Topic Detection and Tracking; Event-based Information Organization*, pages 17–30. Kluwer Academic Publishers, 2002.

[15] F. Jelinek, R. L. Mercer, L. R. Bahl, and J. K. Baker. Perplexity - a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, 62(1):63, 1977.

[16] H. Jin, R. Schwartz, S. Sista, and F. Walls. Topic tracking for radio, TV broadcast, and newswire. In *DARPA Broadcast News Workshop*, pages 199–204, 1999.

[17] T. Leek, R. Schwartz, and S. Sista. Probabilistic approaches to Topic Detection and Tracking. In J. Allan, editor, *Topic Detection and Tracking; Event-based Information Organization*, pages 67–85. Kluwer Academic Publishers, 2002.

[18] J. Lin, R. Snow, and W. Morgan. Smoothing techniques for adaptive online language models: topic tracking in tweet streams. In *SIGKDD '11*, pages 422–429. ACM, 2011.

[19] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma. Terrier: A high performance and scalable information retrieval platform. In *OSIR '06*, volume 2006, pages 18–25, 2006.

[20] M. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[21] G. Smith. *Video scene detection using closed caption text*. Virginia Commonwealth University, 2009. available at `https://digarchive.library.vcu.edu/handle/10156/2649`.

[22] N. Stokes and J. Carthy. Combining semantic and syntactic document classifiers to improve first story detection. In *SIGIR '01*, pages 424–425, 2001.

[23] S. Strassel, J. Kong, and D. Graff. TDT4 Multilingual Text and Annotations. *Linguistic Data Consortium, Philadelphia*, 2005.

[24] K. Zhang, J. Zi, and L. Wu. New event detection based on indexing-tree and named entity. In *SIGIR '07*, pages 215–222, 2007.