Extracting Significant Time Varying Features from Text

Russell Swan and James Allan

Center for Intelligent Information Retrieval
Department of Computer Science
University of Massachusetts
Amherst, Massachusetts USA
{swan,allan}@cs.umass.edu

Abstract We propose a simple statistical model for the frequency of occurrence of features in a stream of text. Adoption of this model allows us to use classical significance tests to filter the stream for interesting events. We tested the model by building a system and running it on a news corpus. By a subjective evaluation, the system worked remarkably well: almost all of the groups of identified tokens corresponded to news stories and were appropriately placed in time. A preliminary objective evaluation was also used to measure the quality of the system and it showed some of the weaknesses and the power of our approach.

1 Introduction

We are interested in information organization and exploration for supporting human decision making. Much information comes in the form of streams, where a stream is a collection of tokens arriving in a fixed order, with each token having a time stamp. Examples of data that are in the form of streams are e-mail, Usenet postings, news corpora, financial and corporate data, and scientific knowledge. Recently there has been an increased study of interfaces for accessing information that contains a time component[3, 7, 10, 5]. Most of these systems were developed for displaying time based information stored in databases. Databases are in many ways easier to work with than is free text, as the data in a database tend to reside in fields with known semantics, whereas in free text the data must be extracted.

Much information exists in the form of stored streams, where time tags exist with known semantics, but the general content is free text. Here, the problem of extracting information and interacting with it is more open ended than in database interaction. A news corpus is a good example of this kind of data source, and is also a classic source of data for Information Retrieval systems. A news corpus has time tags associated with all documents, and the time tags are a form of meta-data that allow novel search, organization, and interaction methods.

Metadata, when available, allows a more effective system to be built. Frequently information from news organizations comes with additional manually assigned metadata. For example, Reuters supplies a list of keywords for each story, where the keywords are selected from a hierarchical list established for that purpose (the KDT system[5], which has many similarities to our system, analyzed the pattern of occurrence of the assigned keywords). Other types of metadata associated with news corpora are section headings (business, entertainment, local, and lifestyle sections), reporter's names (Gina Kolata of the New York Times covers science stories), and tags indicating story importance (front page news, or lead story). This type of metadata usually incorporates some domain knowledge, which can improve effectiveness at the cost of generality. However, we are more interested in general approaches that make as few assumptions about the data as possible, and in discovering how effective we can make a system that assumes time tags are the only metadata available. This should allow us to build a system that is very easy to retarget; for example, our system, built and tested on a news corpus, should also show reasonable effectiveness on a mail list archive.

When a person first encounters a new database, an obvious question is "What is in there?" For a news corpus, a first question might be "What significant events happened during the time frame covered?" We propose a system that performs a statistical analysis on the contents of the corpus and how they vary over time, and uses this system to identify the most distinctive aspects of a collection. In the case of a news corpus, this system is designed to answer the questions, "What were the major stories? What happened?"

In Section 2 we describe the statistical model, our corpus, our system, and show a brief evaluation of our results. In Section 3 we propose a more formal evaluation, and show those results. Section 4 details our conclusions and directions for future work.

2 System

We are investigating whether elementary statistical techniques can be used to select interesting features from corpora automatically. In this section we develop a simple statistical model for the appearance of tokens in a stream. From this model we select a statistical test to determine if the appearance of a specified token is significant. We show these results using a system running on a small news corpus to validate our model, and evaluate our system subjectively. In Section 3 we propose and examine a more formal evaluation.

	w_0	$\overline{w_0}$
$t = t_0$	A	B
$t \neq t_0$	C	D

Table 1: 2 × 2 contingency table

2.1 Model

For a collection of e-mail, or a news corpus, the tokens are the words in the messages or articles, and the time stamp can be specified as the day the message arrived, or the day and hour, or just the month. A simple statistic for discrete events – the presence or absence of a specified token – is the number of tokens arriving during a specified time interval. The model for the arrival of these tokens is a random process with an unknown distribution. (Some research in IR supports the use of a multi Poisson model for these tokens[8], but we are not concerned with the actual distribution here.)

With the model that tokens are emitted by random processes, we assume two hypotheses as defaults. The assumptions are 1: the random processes generating tokens are stationary, meaning that they do not vary over time, and 2: the random processes for any pair of tokens are independent.

If the process producing token w_0 is stationary, then for an arbitrary time period t_0 the probability of seeing the token is the same as the probability of seeing the token at other times. Specifically, looking at the number of times we see w_0 at t_0 (A in Table 1), the number of times we do not see w_0 at t_0 (B in Table 1), the number of times we see w_0 at $t \neq t_0$ (C in Table 1), and the number of times we do not see w_0 at $t \neq t_0$ (D in Table 1), gives a 2 × 2 contingency table. A 2 × 2 contingency table of count data is modeled by a χ^2 distribution with one degree of freedom.

The assumption that two features w_i and w_j have independent distributions implies that $P(w_i) = P(w_i|w_j)$. The resulting counts also form a 2 × 2 contingency table, and are also modeled by a χ^2 distribution with one degree of freedom, with A being the number of times that w_i and w_j occur together, B being the number of times that w_i occurs without w_j, C being the number of times that w_j occurs without w_i, and D being the number of times that neither occur.

For a χ^2 value of 7.879, there is a 0.005 probability that a feature from a stationary process would be identified as not being stationary, or that two independent features would be identified as not being independent. For a corpus containing on the order of 1000 distinct features occurring daily, we would expect on the average five spurious hits/day. If we restrict our attention to multi-day events, where a feature must have occurred by a statistically significant amount for two consecutive days, the probability of a single random token passing the test two days in a row is 2.5×10^{-5}.

2.2 Corpus and Features

Our test material was extracted from the Topic Detection and Tracking (TDT) pilot study's corpus[2]. That material consists of manually transcribed news articles from CNN broadcast news and Reuters newswire from July 1, 1994, to June 30, 1995. We ran each article through a shallow parser[12] to find all noun-phrases, and also through a named entity extraction system called Badger IE[6]. Badger parsed the text to find locations, organizations, and names of people. The corpus was enhanced to include these named entities with markups. Unfortunately, the extraction system was built and tuned for another collection and it was too fragile to work well on all of our test corpus.

The original TDT corpus included 15,683 news stories. Failures of the Badger system forced us to use a subset of those stories. Specifically, we used stories 9001 through 15683—that is, 6683 stories over 175 days spanning January 7 through June 30, 1995. This reduction results in a corpus that would be too small for traditional Information Retrieval effectiveness experiments, but that is fine for our early investigations in time-based organization.

We performed a simple name normalization on the Badger output, as that functionality was missing from the system. The name normalization consisted of conflating all person names with the same last name, and replacing it with the most frequent occurrence, so that for example, the names "McVeigh", "Tim McVeigh", and "Timothy James McVeigh" were all replaced with "Timothy McVeigh". This list of substitutions was automatically generated and hand checked with the substitution being allowed if there were no ambiguity. After the name normalization there were 18421 unique features recognized in the system. Of these, 2030 appeared in five or more documents.

We performed no normalizations or stemming on the extracted noun phrases. The only post processing step after the extraction of the noun phrases was switching all characters to lower case. There were 244,434 distinct noun phrases in the corpus, of which 19,509 occurred in five or more documents.

2.3 Feature Selection

We were interested in determining whether the appearance of a feature is random or not. We did this by performing a χ^2 test for every feature on every set of documents comprising a day. In order to perform the χ^2 test it was first necessary to define what we were taking as samples, and what we were defining as an occurrence. Some choices for what constitutes a sample might be every token, every group of 100 contiguous tokens, or every document. We defined our samples as documents, and we chose for an occurrence any document that had one or more occurrences of that feature. This statistic is referred to as *df* (document frequency). Another statistic frequently used is *term frequency* (*tf*), which is the number of occurrences of a term in a document. *Tf* is monotonically related to the importance of a term, but most models in IR are nonlinear—e.g., $tf/(tf+1)$. For our simple statistical model *df* is easier to implement and more sensible than *tf*. For sample, we use the number of documents that occur in the time of interest.

We only calculate statistics for terms for which $df > 4$. Very infrequent terms are difficult to estimate accurately. On large corpora this has little effect, but on our small corpus we expect that this may cause us to miss some relevant stories.

For each term, and each date, we calculate our χ^2 value. If it is above 7.879 ($p < 0.005$) we conclude that the term's appearance on that day is significant, and begin tracking it. We assemble the largest contiguous block of days we can, where for every day in the block the occurrence is significant. For our multi-day system, we only report features that are significant for spans of greater

Run	Feature	Min Days	# Features	# Used	Sig Features	Sig Stories
1.	Named Entity	2	18421	2030	77	28
2.	Named Entity	1	18421	2030	560	254
3.	Noun Phrase	2	244434	19509	564	115
4.	Noun Phrase	1	244434	19509	5556	2116

Table 2: System runs. The Feature column shows which feature our system reviewed, Badger extracted named entities or noun phrases. Min days shows the minimum number of consecutive days a feature had to appear above our threshold to be reported. # Features is the total number of distinct features in the corpus, and # Used is the total number with $df>4$. Significant features shows how many features were flagged, and significant stories shows how many stories were formed from the features.

Feature	Date Range
Oklahoma City (loc)	April 20 - April 29
Kobe (loc)	Jan 16 - Jan 20
Oklahoma (loc)	April 20 - April 27
FBI (org)	April 20 - April 27
Timothy McVeigh (pers)	April 21 - April 28
NATO (org)	June 2 - June 5
John Doe (pers)	April 21 - April 27
Japan (loc)	Jan 16 - Jan 20
Osaka (loc)	Jan 16 - Jan 18
NATO (org)	May 25 - May 27

Table 3: Top 10 named entities by χ^2 value

Feature	Date Range
oklahoma	April 20 - April 29
oklahoma city	April 20 - April 29
f-16	June 2 - June 5
kobe	Jan 16 - Jan 20
bosnia	May 25 - June 8
bombing	April 20 - April 29
quake	Jan 16 - Jan 20
bosnian serbs	May 25 - June 8
serbs	May 24 - June 6
bosnian	May 25 - May 26

Table 4: Top 10 noun phrase features by χ^2 value

than one day. For the single day system, we report all features that are judged significant. Since our corpus contains on the order of 1000 distinct features per day, we expect spurious results from the test on the single day results, but for the multi-day results spurious features are unlikely.

For example, the χ^2 values for the feature *Oklahoma City*, starting with April 17, are

April 17th	18th	19th	20th	21st
1.38	2.31	1.10	*617.96*	*170.49*
22nd	23rd	24th	25th	26th
208.85	*49.04*	*81.06*	*112.82*	*128.33*
27th	28th	29th	30th	May 1st
95.01	*83.85*	*21.11*	7.26	0.58
2nd				
17.79				

On April 20, the feature has a value of 617.96, above the threshold, so we begin tracking it. It stays above the threshold until April 30, when the value drops to 7.26, so we treat this entire block as being a feature from a story. On May 2, the value goes above the threshold again, but we treat this as a new story.

The ten most significant features as calculated by the Badger based system are shown in Table 3. The most significant features as calculated by the noun phrase system are shown in Table 4. The single day and the multi-day systems had the same top ten features. The single day system, as expected, tagged more features, as shown in Table 2.

Story	Date Range
Oklahoma City Bombing	April 20 - April 29
Earthquake in Kobe, Japan	Jan 16 - Jan 20
F-16 shot down over Bosnia	June 2 - June 5
NATO forces in Bosnia	May 25 - May 27
Flooding in California	Jan 10 - Jan 11
NATO forces in Bosnia	May 29 - May 31
Senate debates Balanced Budget	Feb 28 - Mar 2
Russia/US Summit	May 6 - May 10
Two Americans Sentenced in Iraq	Mar 25 - Mar 27
Henry Foster rejected by Senate as Surgeon General	June 21 - June 22

Table 5: Top 10 stories as calculated by named entity statistics (labels manually assigned)

Story	Date Range
Oklahoma City Bombing	April 20 - April 29
F-16 Shot down in Bosnia	June 2 - June 5
Earthquake in Kobe, Japan	Jan 16 - Jan 20
NATO air strikes in Bosnia	June 2 - June 3
Senate debates Balanced Budget	Feb 28 - Mar 2
Flooding in California	Jan 10 - Jan 11
???(*march, kuwait*)	Mar 21 - March 30
Scott O'Grady rescued	June 7 - June 10
???(*june, saturday*)	June 8 - June 17
U.N. Peacekeepers in Bosnia	June 5 - June 7

Table 6: Top 10 stories found by multiple day noun phrase features

2.4 Story Generation

The features (and their associated date ranges) that we have identified are produced by news stories and events. For example, Table 3 includes *Oklahoma City*, *Oklahoma*, *FBI*, *Timothy McVeigh*, and *John Doe*, all of which are names from the same story (note the nearly identical date ranges). Table 4 shows a similar effect for noun phrases. For a given news story there are usually multiple fea-

Story	Date Range
Oklahoma City Bombing	April 20 - April 29
F-16 Shot down in Bosnia	June 2 - June 5
Earthquake in Kobe, Japan	Jan 16 - Jan 20
NATO air strikes in Bosnia	June 2 - June 3
Senate debates Balanced Budget	Feb 28 - Mar 2
Flooding in California	Jan 10 - Jan 11
???(*march, saturday, plainfield, indiana youth center, ...*)	Mar 21 - March 30
Middle Eastern Terrorists - (Oklahoma City Bombing)	April 20 - April 20
Scott O'Grady rescued	June 7 - June 10
???(*june, saturday, reuter, beijing, ...*)	June 8 - June 17

Table 7: Top 10 stories found by single day noun phrase features

tures that are associated with it. Grouping those features together reduces the total number of stories that must be comprehended, and also makes those stories easier to identify.

To group the stories, we first sort the features on significance. For each feature not grouped into a story we compare its date range with that of lower ranked unassigned features. If there is any overlap in the date range, we test the default assumption that these features are independent over the time span in question. If our test shows that independence is statistically unlikely, we mark the terms as related. For example, consider the named entities of Table 3. For *Oklahoma City* we do not consider merging it with *Kobe* since the dates do not overlap. However, *Oklahoma* does overlap, so we consider the χ^2 value for the pair of terms for that date range. That value is 114.88, well above our threshold, so they are merged. The next feature that has a date overlap, *FBI*, shows a value of 104.00, so it too is merged. We continue until no more candidates for merging exist.

The top ten stories as calculated by the Badger based system are shown in Table 5, with both the single- and the multi-day versions finding the same ten stories. The single day and multi-day versions of the noun phrase based system generated different lists of the top ten stories, and these are given in Tables 6 and 7.

In general, the grouped terms are of high quality and clearly show important features of the stories. The top story (Oklahoma City) as found by the Badger based system was based on the following terms:

> Oklahoma City, Oklahoma, FBI, Timothy McVeigh, John Doe, Justice Department, Michigan, Natalie, Nichols, Terry Nichols, Bernie, Justice Department, Judy Woodruff, Bernard Shaw.

The second story (Kobe earthquake) had the following features:

> Kobe, Japan, Osaka, CNN, Mike Chinoy, Tokyo, Tom Mintier, Andrea Koppel.

(The names of CNN reporters tend to appear with the story they report on.) The same stories, as found through noun phrases, are

> oklahoma, oklahoma city, bombing, building, mcveigh, fbi, john doe, doe, timothy mcveigh, timothy, city, suspect, suspects, explosion, federal building, blast, federal building, terrorism, justice department, debris, bomb oklahoma city bombing, city bombing, law enforcement, nichols, connection, tragedy, rubble, law enforcement, investigation, investigators, michigan, justice, oklahoma city bombing, city bombing, sort, explosives, enforcement, individuals, authorities, enforcement, james, natalie, middle, agents, truck, search, bodies, scene, buildings, terry, survivors, truck, firearms, terry nichols, rescue workers, federal, search, investigators, tuesday, pit, material witness, bernie, oklahoma bombing, bernard, investigation, grand jury, custody, brothers, social security, law, james, indication, justice department, bomb, something, crime, information, federal, speculation, trade, april, thursday, friday, lives, material, federal office building, judy woodruff, bernard shaw, federal office, office building, shaw, agents, explosion, nichols, judy, terry, pieces, farm, hearing, reuter, justice.
>
> kobe, quake, earthquake, fires, japan, devastation, osaka, buildings, supplies, rubble, blankets, opening statements, damage, hit, missing, relief, cnn, tokyo, mike chinoy, chinoy, emergency, judge lance ito, judge lance, reuter, shelters, magnitude, opening, gas, food, scale, simpson case, toll, tom mintier, mintier, rescue workers, death toll, rescue, disaster, streets, roads, survivors, hour, areas, andrea, criticism, tom, correspondent, lines, area, destruction, koppel, andrea koppel, neighborhoods, port, traffic, problem, death, aftershocks, earthquakes, highways, cities, relatives.

The Badger derived features tend to be of higher quality, and the sets tend to be smaller. The noun phrase based stories are quite verbose, and have some clear errors (*opening statements*, *judge lance ito* and *simpson case* are clearly not part of the Kobe earthquake story), but they also have extremely descriptive phrases scored quite highly (*bombing* and *earthquake*, respectively).

The entries labeled "???" in Tables 6 and 7 correspond to "stories" found by the system that we were unable to determine what the story might be. The terms associated with each "story" are:

> march, kuwait
>
> june, saturday;
>
> march, saturday, plainfield, indiana youth center, indiana youth, don king, indianapolis, jeff flock, flock, spokesman, beauty pageant, pageant, incarceration, bodyguards, estai, miss, international waters, fishing, airlines, tuesday;
>
> june, saturday, reuter, beijing, ambassador, polls, monday night, adolfo, tens.

2.5 Discussion

Several errors appear in the collection of stories. The named entity derived stories show three different stories about NATO involvement in Bosnia, but this was instead a single long running story. The requirement we imposed

that a feature had to appear by an amount greater than the threshold for every day in the range caused these stories to stop and then restart as soon as a single day happened when the coverage of the story dropped. For stories of this nature, we clearly need to relax that requirement.

Another problem occurs when the main focus of a story drifts over time, and the system interprets this as being separate stories. For example, when the F-16 was shot down over Bosnia, it received extensive coverage, but the pilot's name (Scott O'Grady) was not released until after he was rescued. Once he was rescued, the style of coverage changed, and his name appeared where it had not before. The selections of terms were sufficiently different that these stories were interpreted as unrelated. A similar phenomenon occurred with the stories on the first day of reporting of the Oklahoma City Bombing, where officials theorized the bombing was the work of a Middle Eastern terrorist group (the eighth story in Table 7). When an American suspect was apprehended, the terms used in the coverage changed so significantly that the stories were interpreted as unrelated. The ongoing research in Topic Detection and Tracking[9] should be helpful in resolving this problem, since it is concerned with gathering stories on the same news topic.

The noun phrases selected formed a very large set, and contain many spurious occurrences. For example, every day of the week except Wednesday appeared as a significant term at some point in the analysis, even though these terms offer no clues as to what a story is about. The noun phrases also assembled "stories" that were not understandable, even after viewing the documents where the phrases occurred. This is a problem similar to that which arises in document clustering where the polythetic nature of the cluster makes it difficult to describe with a list of keywords. Other types of organization – for example, subsumption hierarchies[11] – may be useful for resolving this problem.

From these results it can be seen that the named entity extracted features reduced the set of features to be considered by a greater amount than the noun phrases. The list of the top ten stories were similar for all methods, but the quality of the top ten was greater for the named entity features than for the noun phrases. The noun phrases clearly identified spurious occurrences, and the single day noun phrases had many spurious occurrences. The multiple day named entity features reduced the list of stories to a very small set, all of which had high correspondence to significant news events.

3 Proposed Evaluation

Information Retrieval systems are usually judged by comparing a system's results (on a fixed set of queries on a fixed corpus) with documents judged relevant or not relevant by human assessors. A subjective evaluation shows that the stories found by our system tend to be of high quality for an automatic system–i.e., most of the produced "stories" are reasonable.

An objective evaluation calls for a set of judgments by persons other than the experimenters. We are not aware of any corpora with judgments for tasks similar to what we have done here—finding the top news stories from a corpus. Fortunately, organizations exist whose function it is to summarize and provide top news stories for specified periods. We decided to use the Year in Review for 1995 as given in Facts on File[1] as our judged set. Facts on File Year in Review is a text narrative of the major news stories of the year, with a description ranging in length from one sentence to five sentences for each of the stories. We hired two undergraduates to take the list as given in Facts on File, and reduce it to a machine readable form, where for each story a date range was given, significant names that might be found by named entity were listed, and significant noun phrases were listed. An example is given in Table 8. Each student produced her list independently and they adjudicated them to resolve their differences. We used the most restrictive combination, removing items found by only one person. We used the named entity extracted features from the corpus as a dictionary to verify spelling and forms of names. We then selected only those stories that occurred in the same date range as our corpus, and had at least one name or location associated with them. This resulted in a list of 24 stories, which are listed in Table 9.

3.1 Evaluation Results

Judging stories from Facts on File as relevant, and all stories not listed as not relevant, we performed precision-recall evaluations for all four of the runs, first judging a feature as relevant if the feature was listed in the Facts on File and the dates overlapped, then judging a story as relevant if the dates of the stories overlapped and there was at least one feature in common between the derived story and the judged story. (The dates of the system derived stories were expanded by one day in each direction. The system tagged dates are the dates of the news coverage. Frequently news coverage is the day after an event happens. News stories that are reported from Asia are sometimes reported the day before they happen, due to the time zones crossing the International Date Line.) The matches were then reviewed to verify that they were not caused by spurious random occurrences.

3.1.1 Feature-level Results

We expected poor correspondence between the features, due to vocabulary differences and different word selection, and we were not disappointed. The judged stories contained 62 named features that could have been captured by Badger. Badger generated 77 multi-day features, of which nine were in the relevant set (12% precision, 14% recall). Badger generated 560 single day features with an overlap of 13 (2% precision, 21% recall). The results were worse for the noun phrases. There were 128 noun phrases in the story descriptions. Our system generated 560 multi-day significant noun phrase features, of which only five overlapped (1% precision, 4% recall). While these numbers are very poor, we did not expect much better, due to vocabulary mismatch, lack of stemming in the noun phrases, and the fact that we were comparing a few selected features thought to be descriptive of a story.

3.1.2 Story-level Results

We expected better results from the story matching, and here we were disappointed. The multi-day named entity features generated 28 stories, of which only seven overlapped (stories 8, 11, 13, 16, 19, 22 and 23 were labeled as significant, and lasted more than a day)(25% precision, 29% recall). Four additional stories were detected with named entities, but these stories were in the news for

An earthquake measuring 7.2 on the Richter scale Jan. 17 hit Kobe, Japan, causing more than 5,000 deaths	Large Earthquake hits Japan, more than 5,000 dead DATE 1/17/95 LOCATION Kobe LOCATION Japan Richter scale earthquake

Table 8: Facts on File raw stories and machine readable versions. The left box shows the story as it appeared in Facts on File. The right box shows the format generated by our students, where the top line is a tag to be read by the researcher and ignored by the machine, and the remainder of the information (dates, named entities, and noun phrases) represent important information about the story.

	Story	Date
1.	Egyptian President Mubarak escapes assassination in Ethiopia	June 26
2.	Clinton signs executive order suspending trade with Iran	May 8
3.	Croatian troops retake Western Slavonian and Krajina regions from rebel Serbs	May 1 - May 3
4.	British and Irish Prime Ministers unveil Ulster plan for future talks on Northern Ireland	Feb 22
5.	France elects Jacques Chirac President	May 7
6.	France's President, Jacques Chirac, selects Alain Juppe as premier	May 17
7.	Britain's oldest investment bank, Barings PLC, collapses	Feb 26
8.	Mexico receives international loans	Jan 31
9.	Columbian authorities arrest leaders of Cali drug cartel	June 9
10.	Haiti's Lavalas party wins majority in elections	June 25
11.	Large Earthquake hits Japan, more than 5,000 dead	Jan 17
12.	Cult leader arrested for nerve gas attack on Tokyo's subways	May 16
13.	Fatal nerve-gas attack on Tokyo's subways	March 20
14.	John Howard elected leader of Australia's Liberal Party	Jan 30
15.	Taiwan President Lee Teng-hui visits U.S., angers China	June 7 - June 10
16.	Ebola virus outbreak in Zaire	May 11
17.	U.N. peace-keepers mission in Somalia ends	March 3
18.	Arkansas Governor Jim Guy Tucker (D) indicted on fraud charges in connection with the Whitewater affair	June 7
19.	Dr. Henry W. Foster Jr.'s surgeon general nomination derailed by Senate Republicans	June 21 - June 22
20.	Senate confirms John M. Deutch director of central intelligence	May 9
21.	U.S. Supreme Court blocks minority districting	June 29
22.	Bomb explodes outside Oklahoma City federal building, Timothy McVeigh and Terry Nichols arrested and charged	April 19
23.	Floods ravage California	Jan
24.	Floods ravage California	March

Table 9: Facts on File stories

only one day (stories 1, 5, 12, and 24). They were captured by our single day system, at the expense of a far larger set to evaluate (254 stories as opposed to 28 stories)(4% precision, 46% recall). The noun phrase based system generated far more stories than the Badger based system, (115 instead of 28 multi-day stories, 2116 instead of 254 single day stories) and found one additional story (story 7). The name "Barings" was not normalized, and the several different variants caused our Badger based system to miss this story. If the name had been normalized it would have been found by our single day system, but not by the multi-day system. Twelve of the 24 stories were not flagged. We identified several causes of error that identify why the 12 stories were missed:

- Of the 12 stories missed, four were never mentioned in our corpus (stories 2, 4, 14 and 21), and two were mentioned in only one story (stories 6 and 20).
- One story was briefly covered. Story 18 had two mentions. The most distinctive features of the story (*Tucker*, *Madison Guaranty*) had *df* values below our threshold (3 and 1, respectively) and statistics were not calculated.

Five desired stories were missed by our systems. Stories 3 and 9 were covered in our corpus, but the coverage was spread out over several days (story 9 was written about in one story per day for five consecutive days), and there was never a single day where the coverage appeared as significant. Stories that were covered well in the corpus but missed by our system were stories 10, 15 and 17. These stories were missed because the features that were distinctive about them (Haiti, Taiwan) were frequently in the news, and the occurrence of those features on that specific day was not that different than their occurrence on any other day.

Interestingly, the Badger based multi-day system found two stories related to stories 10 and 17 that were *not* reported in Facts on File. (1) President Clinton visiting Haiti three months before the election, as U.S. troops were preparing to turn over their duties to the U.N. (2) The U.S. fleet helping in the U.N evacuation from Somalia was flagged as a story on February 26, five days before the last troops left. (Due to our requirement that the date ranges had to overlap within one day, this was counted as a miss, even though both the Facts on File story and our generated story were about the withdrawal from Somalia.) The coverage in Facts on File tends to stress dates when politicians and diplomats did things, whereas CNN coverage tends to be more concerned with what the troops are doing. In addition, the Badger based multi-day system identified several stories that were arguably significant, but not listed in Facts on File. These included the G7 summit meeting in Halifax, Nova Scotia that Boris Yeltsin was invited to (sometimes referred to as "the G7+1 summit").

Of the 24 stories identified in Facts on File, four were never mentioned in our corpus and two were mentioned once. Of the remaining 18, seven were identified by our multi-day Badger based system, four additional stories were identified by our single day system, and one story was caught by our noun phrase system that was missed by our Badger based systems.

4 Conclusions and Future Work

We have devised a simple statistical model and achieved very strong results for a simple, automatic model. The stories generated have been reasonable to people viewing the results.

Named entity features are far more accurate. Noun phrases tend to be a lot noisier, causing very large sets and large numbers of spurious terms. Noun phrases are far more descriptive, and make well identified stories easier to understand, but they are not very valuable in the initial selection of the stories.

Multi-day features are very unlikely to get errors, but the threshold is too high – we miss good stories. Single day features catch the short stories, but also make a lot of errors. (We are performing thousands of comparisons for each day's news, and each comparison has an 0.005 probability of mislabeling a non-significant feature.) A possible method of increasing the recall of our system without hurting the precision too greatly is to have two thresholds, a lower one for multi-day stories, and a very high one ($p < 0.0001$) for single day stories.

Ultimately, we want to find stories automatically and present them in a way that will be sensible to a human. Using named entity features for finding a story, then using the most significant noun phrases for labeling the stories seems promising. Many of these stories change over time, and we want to be able to quickly and succinctly show both the information and how the information changes to users. We are beginning to investigate visualizations for that purpose.

We also intend to further investigate the statistics we focus on. For two terms, t_1 and t_2, we can ask "Are they significantly related?" and "What is the relationship between them?" Relationships between terms are often modeled by the amount of information the presence of t_1 provides about t_2. This is often measured by statistics describing entropy, such as the Kullback-Leibler measure[4]. Entropy based measures provide good indications about the relationship between terms once it is established that terms are related; however, they are poor at determining if terms are related compared to statistics such as χ^2. We have seen good results with using χ^2 for answering the first question, and we will begin investigating other statistics for answering the second question.

Our methodology for evaluation seems reasonable, but clearly needs improvement. The bulk of the work appears to be resolving vocabulary mismatches and aligning the "truth" information with the corpus. The preliminary nature of the evaluation exposes both the power and the weaknesses of our approach and suggests directions to explore.

5 Acknowledgments

We would like to thank David Fisher for his help in getting Badger running, Victor Lavrenko for running Badger and the noun phrase extractor on the TDT corpus and for his general help and comments, David Jensen for reviewing our mathematics and analysis, and Lesley Lam and Rebecca Zambrowski for adapting the Facts on File stories to machine readable form.

This material is based on work supported in part by the National Science Foundation, Library of Congress and Department of Commerce under cooperative agreement number EEC-9209623, and also supported in part by the National Science Foundation under grant number IRI-9619117. Any opinions, findings and conclusions or recommendations expressed in this material are the author(s) and do not necessarily reflect those of the sponsor.

References

[1] *Facts on File, 1996*. Facts on File, New York, 1997.

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

[3] R. B. Allen. Timelines as information system interfaces. In *Proceedings International Symposium on Digital Libraries*, pages 175–180, Tsukuba, Japan, 1995.

[4] Yvonne M. M. Bishop, Stephen E. Feinberg, and Paul W. Holland. *Discrete multivariate analysis: theory and practice*. MIT Press, Cambridge, Massachusetts, 1974.

[5] Ido Dagan and Ronen Feldman. Keyword-based browsing and analysis of large document sets. In *Proceedings of the Symposium on Document Analysis and Information Retrieval (SDAIR-96)*, Las Vegas, Nevada, 1996.

[6] D. Fisher, S. Soderland, J. McCarthy, F. Feng, and W. Lehnert. Description of the umass systems as used for muc-6. In *Proceedings of the 6th Message Understanding Conference, November, 1995*, pages 127–140, 1996.

[7] Robin L. Kullberg. Dynamic timelines: Visualizing historical information in three dimensions. Master's thesis, Massachusetts Institute of Technology Media Laboratory, 1995.

[8] E. L. Margulis. Modeling documents with multiple poisson distributions. *Information Processing and Management*, 29(2):215–227, 1993.

[9] Ron Papka, James Allan, and Victor Lavrenko. Umass approaches to detection and tracking at TDT2. In *Proceedings of the DARPA Broadcast Workshop*, 1999.

[10] C. Plaisant, B. Milash, A. Rose, S. Widoff, and B. Shneiderman. Life lines: Visualizing personal histories. In *CHI'96 Conference Proceedings*, pages 221–227, Vancouver, BC, Canada, 1996.

[11] M. Sanderson and W. B. Croft. Deriving structure from texts. In *Proceedings of the 22nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR99)*, 1999.

[12] Jinxi Xu, J. Broglio, and W. B. Croft. The design and implementation of a part of speech tagger for english. Technical Report IR-52, Center for Intelligent Information Retrieval, University of Massachusetts, Amherst, 1994.