Modeling Document Dynamics: An Evolutionary Approach

Jahna Otterbacher, Dragomir R. Radev

University of Michigan
Ann Arbor, Michigan, USA
{jahna,radev}@umich.edu

Abstract

News articles about the same event published over time have properties that challenge NLP and IR applications. A cluster of such texts typically exhibits instances of paraphrase and contradiction, as sources update the facts surrounding the story, often due to an ongoing investigation. The current hypothesis is that the stories "evolve" over time, beginning with the first text published on a given topic. This is tested using a phylogenetic approach as well as one based on language modeling. The fit of the evolutionary models is evaluated with respect to how well they facilitate the recovery of chronological relationships between the documents. Over all data clusters, the language modeling approach consistently outperforms the phylogenetics model. However, on manually collected clusters in which the documents are published within short time spans of one another, both have a similar performance, and produce statistically significant results on the document chronology recovery evaluation.

1. Introduction

When an important event happens, large numbers of news sources report on it. In doing so, they draw information from direct participants in the event, eyewitnesses, official reports, copy from the newswire, as well as from each other. As anyone who follows an event can attest, often multiple sources present complementary accounts of the news. Each source has its own reputation, biases, and agenda. In addition to source, news accounts of an event vary over time. Often initial reports turn out to be partially or fully incorrect. It takes a certain amount of time for accounts to stabilize and to be accepted as the ground truth.

In considering how information evolves over time and is expressed through text, we have examined sets of documents on the same story published over time by multiple news agencies, and have found that they exhibit a number of interesting relationships. For example, a given pair of related documents may express some of the same factual information and yet each may contain novel information that the other does not. An example with respect to a single fact is illustrated in Figure 1. The sentences shown were extracted from documents describing the crash of a small plane into a skyscraper, and concern the location from where the plane departed.

In short, following information in a news story over time and across sources is a challenging task due to the dynamic nature of such texts. As facts, beliefs and opinions surrounding an event change, so do the texts that report on them. In other words, such stories can be viewed as "evolving" over time, beginning with the information reported in the first story that makes the news. Currently, we attempt to model these phenomena using a phylogenetic approach. In phylogenetics, the history of a set of species is reconstructed, under the assumption that they evolved from a common ancestor, with genetic mutations occurring at different points in time. The "species" we study are related documents describing the same news story.

In addition, we will test a second approach that is inspired by language modeling. We use a language model generated from the earliest document in each set, to chronologically order the remaining documents. In doing so, we hypothesize that as time goes on and the story changes, the likelihood that the original language model could have generated a later document should decrease. In both experiments, we evaluate the fit of the evolutionary models with respect to their ability to recover the chronological relationships between the documents in a given cluster. Rather than experimenting with a large number of text representation methods within each approach, we have applied the same preprocessing techniques to the texts in the corpus before implementing the models. It is likely that we will be able to improve the performance of both approaches on the chronology recovery task in our future work. However, the goal of the current paper is to evaluate the extent to which multi-document clusters of news articles exhibit evolutionary properties as well as to see which approach, phylogeny or language modeling, is more promising for modeling inter-document dynamics.

2. Related work

2.1. A method for phylogenetic analysis

The Fitch-Margoliash method is used in the biological sciences for constructing a phylogenetic tree for a set of species, based on sequences of amino acids found in their DNA (Fitch and Margoliash, 1967). First, mutation distances are calculated between each pair of species. This distance is the minimum number of sites that would have to be changed in order for one string to mutate into the other. Initially, each of the N species is assigned to its own subset, such that there are N subsets. They are then joined together, starting with those that have the smallest mutation distance between them, such that the number of subsets is reduced by one at each cycle, until all subsets have been joined to the tree.

Because of the manner in which the initial sets are chosen, various phylogenetic trees will result from the different initial assignments. Therefore, it is necessary to test between alternative trees. For each tree, one sums over the distances between each pair of species, resulting in a new distance matrix that can be compared to the original mutation dis-

```
04/18/02 13:17 (CNN)
The plane, en route from Locarno in Switzerland,
to Rome, Italy, smashed into the Pirelli building's
26th floor at 5:50 p.m. (1450 GMT) on Thursday.

04/18/02 13:42 (ABCNews)
The plane was destined for Italy's capital Rome,
but there were conflicting reports as to whether it
had come from Locarno, Switzerland or Sofia, Bulgaria.

04/18/02 13:42 (CNN)
The plane, en route from Locarno in Switzerland,
to Rome, Italy, smashed into the Pirelli building's
26th floor at 5:50 p.m. (1450 GMT) on Thursday.

04/18/02 13:42 (FoxNews)
The plane had taken off from Locarno, Switzerland,
and was heading to Milan's Linate airport,
De Simone said.
```

Figure 1: Dynamic information example.

tances. The "percent deviation" of the reconstructed values in the tree from the original input distances are found by summing the squared percent change for each species. For example, if the original mutation distances between pairs of species are in the upper triangle of the distance matrix, while the new distances according to the candidate tree are in the lower triangle, then for each species pair the original distance is (i, j) and the new distance is (j, i).

$$\text{Percent deviation} = \sum_{i<j} \left(\frac{|(i,j) - (j,i)|^2}{(i,j)}\right) * 100$$

Seeking the statistically optimal phylogenetic tree from the set of all possible trees involves minimizing the percent deviation.

2.2. Phylogenetic trees and text analysis

Bennett and colleagues applied phylogenetic inference algorithms to reconstruct the evolutionary history of 33 chain letters collected between 1980 and 1995 (Bennett et al., 2003). Because the chain letters circulated before the widespread use of email, they proposed that the letters mutated and evolved as generations of receivers photocopied them until no longer legible. At such a point, the next recipient would likely retype the letter, introducing new errors and variations.

The distance metric between each pair of chain letters x and y used in constructing the tree was the amount of information, $d(x, y)$ shared by the pair of letters. Once the distance matrix was computed, the authors used various methods, including Fitch-Margoliash, in constructing phylogenetic trees. The tree was rooted using the letter with the earliest known date. Using the same distance metric, the various methods for constructing the tree yielded similar trees. Once the tree was constructed, the authors were able to explain how the chain letters evolved over time. For example, names of individuals and the dates of different events mentioned in the letter (such as the death of someone who broke the chain) changed at different points in its evolution. In addition, new "genes" often appeared. The resulting tree was almost a perfect phylogeny, as the authors were able to confirm that letters containing the same characteristics were always grouped together.

2.3. Our approach

The current work is inspired by Bennett's research but differs in some important ways. In the chain letters, mutations occurred over time because of letters being recopied by recipients, who might misspell or misinterpret words in the letter when preparing copies to mail out to the next receivers. Alternatively, details of the letters were occasionally changed deliberately. For example, when the letters were first brought to the U.S. from Europe, certain names and titles were changed. In our work, we assume that over time, we will observe mutations in news stories because they reflect events and facts in the real world that are constantly changing.

There are some other interesting nuances in the current problem. For example, while we assume that the texts we observe express the facts in the world, there is rarely only one way to express the same concept or fact in natural language. Therefore, we expect to encounter many instances of paraphrases in our data. At the same time, it is known that journalists use newswire sources and may also copy large parts of previously published news stories in creating an update on a given situation (Clough et al., 2002; Mitchell and West, 1996). Therefore, we will also observe many instances of identical expressions, published by different sources and perhaps even at different points in time.

In our experiments, we attempt to recover the chronological relationships between related documents using two different approaches. In the first approach, we create an unrooted phylogenetic tree for each document cluster, and then reroot each tree at the document in the cluster that has the earliest publication date. Therefore, S1 (Species 1) is at the base of the tree, and we propose that the remaining docu-

ments arise as mutations occur. Once we have our rerooted tree for a cluster of documents, we calculate the distance from the root, S1, to each of the other documents. Our hypothesis is that these distances should correlate well to the chronological ordering of the documents.

We will compare the performance of the phylogenetic document ordering algorithm to that of a second approach based on language modeling. Language modeling has been used extensively in information retrieval for document ranking. In this setting, a document is considered to be relevant to an information query if the language model built from the document assigns a high probability to the query (Ponte and Croft, 1998). More recently, (Kurland and Lee, 2004) used language models for modeling inter-document relationships. In our experiments, we create a language model from the earliest document in each cluster. We then evaluate it on the remaining documents and use its fit to rank them. Our hypothesis is that the model fit should be better for the earlier documents and degrade as time goes on, since as the facts in the story change, new terms and expressions arise.

3. Corpus

Table 1 shows the characteristics of the document clusters used in the experiments. Six clusters were collected manually by the authors, three (Bali bombing, Turkish Air crash and Hamas bombing) were collected automatically from our Web-based news tracking system and 27 clusters were taken from the TREC Novelty Track 2003 and 2004 test sets (Soboroff and Harman, 2003) [1]. They were randomly assigned to the training (15 clusters), development/test (6 clusters) and test data sets (15 clusters), although we did ensure that they were distributed to each data set rather evenly by type.

As can been seen, the Novelty clusters differ from our manually collected clusters in one important way. While the manual clusters were collected over a relatively short time period (e.g. a few days), the Novelty clusters typically contain documents published over a much wider time span. In addition, our manually collected clusters all describe emergency news stories (e.g. plane crashes, fires), while the Novelty clusters include a wide range of topics. For use in the experiments, all texts in the corpus were tokenized, such that all punctuation was removed and all capital letters were made lowercase.

Story	Doc.	Time span	Sources	Data set
Milan plane crash	56	1.5 days	5	train
RI nightclub fire	43	1.5 days	8	train
Iraq bombing	30	1.5 days	10	train
Turkish Air crash	10	6 days	4	train
N4 - EgyptAir crash	25	8 months	3	train
N6 - Unabomber	25	3.5 years	3	train
N8 - Berenson imprisoned treason Peru	25	4.5 years	3	train
N33 - Russian submarine Kursk sinks	25	1 month	3	train
N34 - Glenn Shuttle Discovery	25	1 month	3	train
N42 - JFK Jr. dies	25	1 year	3	train
N43 - Chinese earthquake	25	1 year	2	train
N44 - Plane gondola cable accident	25	1 year	2	train
N51 - General Pinochet arrested	25	10 months	3	train
N64 - Japan nuclear accident	25	1 year	3	train
N87 - Birmingham church bombing	27	4 years	3	train
Columbia shuttle disaster	41	2.5 days	6	devtest
Bali bombing	10	13 days	5	devtest
N7 - Atlanta Olympics bombing	25	3.5 years	2	devtest
N49 - 1998 Nobel peace prize	25	3 months	2	devtest
N53 - Death of James Byrd, Jr.	32	1.5 years	2	devtest
N81 - Matthew Shepard	25	1.5 years	2	devtest
GulfAir plane crash	11	1 month	7	test
Honduras bus hijacking	46	2 days	10	test
Hamas bombing	11	2 days	7	test
N9 - Columbine shooting	25	1 year	3	test
N11 - Hurricane Mitch	25	2 months	2	test
N16 - Kenya embassy bombing	25	1 year	3	test
N37 - Olympic bribery scandal	25	2 years	3	test
N40 - Wen Ho Lee, Los Alamos	25	1 year	3	test
N45 - Slepian abortion murder	25	1.5 years	2	test
N48 - Human genome decoded	25	2 years	3	test
N50 - Balloonist Fossett solo flight	25	1 year	2	test
N59 - Steward plane crash	25	1 year	3	test
N69 - Concorde crash	27	2 months	3	test
N80 - Turkey earthquake	41	4.5 years	2	test
N83 - Marine Osprey	25	5 months	3	test

Table 1: Document clusters used in experiments.

4. Phylogenetics experiments

4.1. Document ordering

We applied the phylogenetic technique on the full text of the documents, as well as on summaries produced from each individual document using various compression rates using the MEAD extractive summarizer (Radev et al., 2004). The intuition behind using summarization is that it might highlight the most salient information in each document, while eliminating some information that might not be important for recovering inter-document relationships. For each run on a given document cluster, we calculated the Levenshtein matrix, or the edit distances between all pairs of documents (at the word level). This was used as our mutation distance in order to construct the phylogenetic trees using the Fitch-Margoliash method. We used the Fitch program (part of the Phylip Inference package) to construct the trees (Felsenstein, 1995).

Since Fitch produces unrooted trees, such that we obtain relative distances between documents, rather than from a common starting point, we rerooted each tree at the earliest sentence in the cluster. Our text dynamics rerooting algorithm is shown in Algorithm 1.

4.2. An example

In this section, we illustrate our methods using a small example cluster of four topically related documents from the Milan training cluster. For illustrative purposes, we have represented each document as one sentence extracted from it, rather than showing the entire text of the document. Each document species is shown with its respective publication

[1] We included Novelty clusters that were labeled as describing events only. We did not include opinion clusters.

```
S1: Italian TV says the crash put a hole in the 25th floor of the
Pirelli building, and that smoke is pouring from the
opening. (04/18/02 12:22, CNN)

S2: Italian TV showed a hole in the side of the Pirelli building with
smoke pouring from the opening. (04/18/02 12:32, CNN)

S3: Italian state television said the crash put a hole in the 25th
floor of the Pirelli building. (04/18/02 12:42, MSNBC)

S4: Italian state television said the crash put a hole in the 25th
floor of the 30-story building. (04/18/02 12:44, FOX)
```

Figure 2: Sample document "species" in chronological order.

Algorithm 1 TD tree rerooting algorithm.

Root tree at S_1
$depth(S_1) = 0$
Initialize stack q of next documents to process
Push S_1 onto q
repeat
  S_i= next element in q
  $seen(S_i) = 1$
  Find depth of S_i in tree
  depth(S_i)=Find_depth(S_i)
until stack q is empty

Function Find_depth(S_i)
for each element a_i in tree **do**
  b_i is element adjacent to a_i and $distance(a_i, b_i) = c_i$
  if a_i=S_i and seen(b_i)=0 **then**
    Push b_i onto q
    depth(b_i)=c_i + depth(S_i)
    Return depth(b_i)
  end if
  if b_i=S_i and seen(a_i)=0 **then**
    Push a_i onto q
    depth(a_i)=c_i + depth(S_i)
    Return depth(a_i)
  end if
end for

date, time stamp and source in Figure 2.
First, the Levenshtein matrix is calculated, yielding the distance matrix for Fitch. The distance matrix for the above example is shown in Figure 3. Each entry (i, j) in the matrix shows the word-level edit distance between document i and j. Note that the Levenshtein matrix is also symmetric with zeros along the diagonal.

	S1	S2	S3	S4
S1	0	10	12	13
S2	10	0	15	16
S3	12	15	0	1
S4	13	16	1	0

Figure 3: Levenshtein matrix for 4 input document species.

Once the best fitting evolutionary tree is found by the Fitch-Margoliash method, it is then rerooted at the earliest document in the cluster. The unrooted tree (output of Fitch) for the example is shown in Figure 4. Note that the tree shows both the document species as well as internal nodes, intermediate points at which a mutations occur. The nodes and species are shown with their respective distances from node I_1, an arbitrary point. The corresponding rerooted tree is shown in Figure 5. Here, the distances shown are from the given node or species to S1, the root. To obtain these distances, the tree is traversed from the root out. The system ranking is then determined with respect to the distances, with species closer to the root having higher ranks. The ranks correspond to the chronological ordering of the document species. To evaluate, the system rankings are compared to the actual chronological ordering of the documents. Figure 6 illustrates this process.

5. Language modeling experiments

As previously mentioned, for each document cluster, a language model was built from the earliest document in the set. More specifically, a simple trigram backoff language model with Good Turing discounting was created and evaluated against every other document in the cluster using the CMU-Cambridge toolkit (Clarkson and Rosenfeld, 1997). Since the first document in a cluster typically had a much smaller vocabulary than latter documents, we used the out-of-vocabulary (OOV) rates as well as the backoff event information rather than model perplexity in order to assess the fit with respect to each document in the cluster. We hypothesized that for documents published later on, the OOV rate should be greater. Likewise, we expect to see more backoff events, such that the trigram-hit ratios should be smaller, and unigram-hit ratios larger, as compared to earlier documents. There were three experiments per cluster: one in which documents were ordered by OOV, by unigram-hit ratio and by trigram-hit ratio (ranked in reverse order). We then compared the system orderings to the true orderings in the same manner as in the phylogenetic experiments.

6. Experimental results

6.1. Evaluation method

For each cluster and system ordering, the Kendall rank-order correlation coefficient was calculated (Siegel and

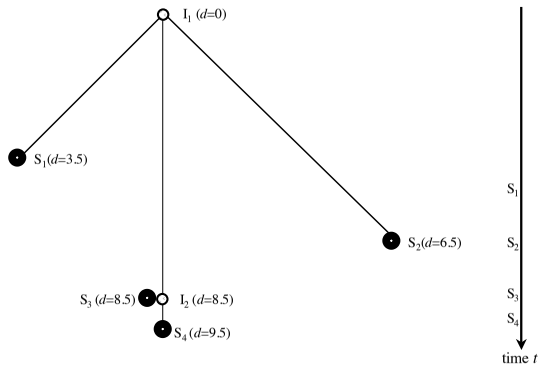

Figure 4: Unrooted tree.

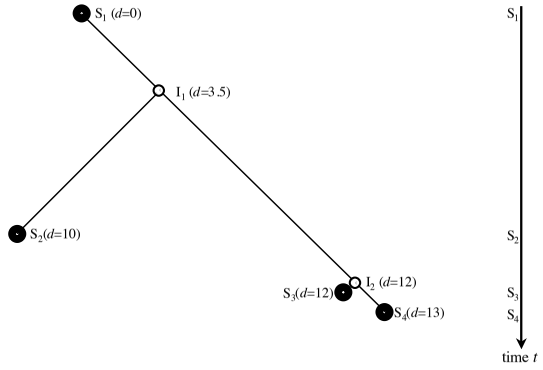

Figure 5: Tree rooted at Species 1 (S1).

Castellan, 1988). Kendall's τ quantifies the extent to which the rankings assigned by the system are correlated to the actual rankings: $\tau = \frac{2*(n_a - n_d)}{N(N-1)}$, where n_a is the number of agreements, n_d is the number of disagreements and N is the number of ranked documents. In the case of tied ranks, there is an adjusting factor in the denominator, such that that penalty is less for a disagreement between the system and the actual ranks.

Essentially, τ is the ratio of the difference between the number of partial ranks in agreement and those in disagreement between the system and the actual rankings to the maximum possible total. Therefore, a τ of 1 indicates that the ranks assigned by the system agree perfectly with the true ranks. Figure 7 illustrates the calculation of τ for the set of example document species.

Comparing the partial rankings of the system to the actual rankings, there are 6 in agreement and none in disagreement. Therefore, $\tau = \frac{2*(6-0)}{4*(4-1)} = 1$.

The p-value for a τ of 1 when N=4 is 0.025. The interpretation of this value is that if we repeatedly draw a sample of four documents from the population of documents related to the Milan story, then under the null hypothesis that the rankings assigned by our algorithm and the actual rankings are uncorrelated, the probability of finding a τ=1 (or a more extreme value of the test statistic) is 0.025. Currently, we will use a significance level of 0.10 for reporting our experimental results.

System	Actual
S1 > S2	S1 > S2
S1 > S3	S1 > S3
S1 > S4	S1 > S4
S2 > S3	S2 > S3
S2 > S4	S2 > S4
S3 > S4	S3 > S4

Figure 7: Comparing partial rank orderings for calculating τ.

6.2. Training phase

In the training phase, we evaluated 11 document ordering mechanisms on the 15 training clusters. We implemented the phylogenetic algorithm on the full text of the documents, as well as on the document summaries at lengths of 1, 2, 3, 4, 5, 6 and 8 sentences. We also evaluated document ordering using the three language modeling approaches (based on trigram-hit and unigram-hit in the back-off model, and OOV as previously discussed). The median Kendall's τ over the 15 document clusters, and the number of clusters on which τ was statistically significant are

Document species	Distance from root	System rank	Actual rank
S1	0	1	1
S2	10	2	2
S3	12	3	3
S4	13	4	4

Figure 6: Chronological ordering of the input documents.

	Med. τ	# Sig.
Full doc	0.16	8/15
Summ-1	0.13	6
Summ-2	0.12	5
Summ-3	0.13	6
Summ-4	0.16	6
Summ-5	0.17	6
Summ-6	0.09	6
Summ-8	0.12	6
3gram	0.17	7
1gram	0.21	11
OOV	0.28	13

Table 2: Median τ and the number of data clusters with a significant result.

	Med. τ	# Sig.
Summ-5	0.05	3/11
1gram	0.20	8/11
OOV	0.19	8/11

Table 3: Median τ and the number of clusters with a significant result for the 11 Novelty training clusters.

shown in Table 2. Over all clusters, the language modeling OOV approach performed the best, having a median τ of 0.28. In addition, for 13 of 15 training clusters, the results were statistically significant.

The best run for the phylogenetic approach was the one which calculated the edit distance between each document species based on the 5-sentence summary of each document. Tables 3 and 4 show the comparison of this approach against the two best language modeling approaches (1gram and OOV) on the 11 Novelty data clusters and the 3 manually-created clusters, respectively. As mentioned in Section 3., the manual clusters differ from the Novelty clusters not only in that all discuss emergency news topics (that are likely to report changes rapidly over time) but also in that the publication times of the documents are relatively closer together. Here we can see that on the manual clusters, all three methods yield statistically significant results on all three manual clusters. However, for the Novelty clusters, 1gram and OOV perform much better than the phylogenetic technique.

6.3. Development/test phase

In the development/test phase, we evaluated the top two language modeling approaches (1gram and OOV) as well as the best two phylogenetic techniques (Summ-4 and Summ-5) in order to distinguish them further in terms of performance. Table 5 shows the τ for each of the six development/test clusters as well as the median over all clusters and the number of significant orderings. In this set, only one cluster, which describes the Columbia shuttle disaster, is a manually-created cluster and as expected, all four techniques achieve a statistically significant result on ordering the 41 documents in the cluster. However, we again observe some poor performances on the Novelty clusters. In particular, Summ-4 achieves a τ of only 0.04 on clusters N53 and N81. Given its lower median τ as well as having a significant performance on only half of the clusters, we eliminate Summ-4 and evaluate the remaining three techniques on the unseen test data set.

	Med. τ	# Sig.
Summ-5	0.32	3/3
1gram	0.42	3/3
OOV	0.26	3/3

Table 4: Median τ and the number of clusters with a significant result for the 3 manual training clusters.

	Med. τ	# Sig.
Summ-5	0.15	5/15
1gram	0.14	6/15
OOV	0.22	9/15

Table 6: Median τ and the number of clusters with a significant result for 15 test clusters.

6.4. Test phase

The performance of the three remaining techniques is shown in Table 6. The technique that orders documents with respect to their OOV rate when evaluated against the language model created by the earliest document in the set outperformed the other two methods. In particular, the OOV technique achieved a statistically significant Kendall's τ on 9 of the 15 unseen test clusters.

7. Conclusions

While over all data clusters, the OOV technique outperformed all others, we have also seen that in general, we achieved better results on the manually-collected document sets as compared to the Novelty clusters. Table 7 shows the performance of the OOV (language model) and Summ-5 (phylogenetic) techniques the six manual clusters over all data sets. To contrast, over all 27 Novelty clusters in our corpus, the median τ for the OOV and Summ-5 techniques was 0.22 and 0.17, respectively. Therefore, one conclusion from our experiments is that the evolutionary models that we have proposed and implemented fit the manual clusters rather well. As previously mentioned, these clusters were collected over shorter periods of time from Web-based

Cluster	OOV	1gram	Summ-4	Summ-5
Columbia shuttle	0.56	0.52	0.46	0.48
Bali bombing	0.20	0.24	0.51	0.29
N7 - Olympics bombing	0.32	0.27	0.15	0.24
N49 - Nobel prize	0	0.29	0.25	0.31
N53 - Death of J. Byrd	0.21	0.27	0.04	0.20
N81 - Matthew Shepard	0.35	0.23	0.04	0.19
Med. τ	0.26	0.27	0.20	0.26
# Sig.	4/6	5/6	3/6	5/6

Table 5: Individual cluster τ, and median τ and significance for all 6 dev/test clusters.

Cluster	OOV	Summ-5
Gulfair plane crash	0.37	0.39
Honduras bus hijacking	0.12	0.17
Columbia shuttle	0.56	0.48
Milan plane crash	0.26	0.33
RI nightclub fire	0.58	0.32
Iraq bombing	0.24	0.17
Med. τ	0.31	0.33
# Sig.	5/6	6/6

Table 7: Performance over all 6 manually-created clusters.

news sources. In addition, we tried to collect as many documents as possible that were published over time describing the given subject, which was an emergency situation.

To contrast, the Novelty cluster topics are more varied and as can be seen in Table 1, the publication time spans are typically larger (e.g. over months or years) rather than over days, as in our manual clusters. It is obvious that our evolutionary models in general, do not fit these types of document clusters as well. In fact, the poorest performances observed in the test data are on Novelty clusters. For example, for the cluster N80 about the Turkey earthquake, which contains 41 documents published over a period of 4.1 years, none of the techniques achieves a statistically significant result. Therefore, we conclude that the evolutionary models are most useful for predicting relationships between documents describing related, breaking news stories and that are published over shorter time intervals.

7.1. Future work

Having shown that clusters of breaking news stories published over time and by different sources have evolutionary properties that can be modeled, we plan to improve and extend our methods for chronology recovery between texts. In particular, we plan to adapt our current techniques to the problem of following changing information in clusters of "evolving" texts, such as the emergency news stories currently studied. Given a set of documents of interest, the goal will be to chronologically order the individual facts they express (rather than the documents themselves). Eventually, we hope to implement this work into a Web-based system that will help users monitor changing news events at the factual level.

8. Acknowledgements

This work was supported in part by NSF Grants IIS 0534323 and BCS 0527513. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation.

9. References

Charles H. Bennett, Ming Li, and Bin Ma. 2003. Chain Letters and Evolutionary Histories. *Scientific American*, pages 76–81, June.

P.R. Clarkson and R. Rosenfeld. 1997. Statistical Language Modeling Using the CMU-Cambridge Toolkit. In *ESCA Eurospeech*.

Paul Clough, Robert Gaizauskas, Scott S.L. Piao, and Yorick Wilks. 2002. Measuring text reuse. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 152–159.

Joseph Felsenstein. 1995. PHYLIP: Phylogeny Inference Package. Technical report, Department of Genome Sciences, University of Washington.

Walter M. Fitch and Emanuel Margoliash. 1967. Construction of Phylogenetic Trees. *Science*, 155(3760):279–284, January.

Oren Kurland and Lillian Lee. 2004. Corpus Structure, Language Models, and Ad Hoc Information Retrieval. In *SIGIR 2004*.

Catherine C. Mitchell and Mark D. West. 1996. *The News Formula: A Concise Guide to News Writing and Reporting*. St. Martin's Press, New York.

Jay M. Ponte and W. Bruce Croft. 1998. A Language Modeling Approach to Information Retrieval. In *SIGIR 1998*.

Dragomir R. Radev, Hongyan Jing, Malgorzata Styś, and Daniel Tam. 2004. Centroid-based summarization of multiple documents. *Information Processing and Management*, 40:919–938, December.

Sidney Siegel and N. John Castellan. 1988. *Nonparametric Statistics for the Behavioral Sciences*. McGraw Hill.

Ian Soboroff and Donna Harman. 2003. Overview of the TREC 2003 Novelty Track. In *Proceedings of the Twelfth Text Retrieval Conference (TREC 2003)*, NIST, Gaithersburg, ML.