Tracking Factual Information in Evolving Text: An Empirical Study

Jahna Otterbacher[1], Siwei Shen[2], Dragomir Radev[1,2], and Yang Ye[3]
[1] School of Information
[2] Department of Electrical Engineering and Computer Science
[2] Department of Linguistics
University of Michigan
Ann Arbor, MI 48109
{jahna,radev}@umich.edu

Abstract

We consider the problem of tracking information over time and across sources in clusters of news stories about emergency events. While previous approaches have focused on finding new information at the document and sentence levels (e.g. TDT FSD and the TREC Novelty track, respectively), we are interested in following information at the factual level. Specifically, we propose to develop a "fact tracking" system, that when given a user's factual question of interest, returns a matrix displaying the extracted answers by time and source. As a first step towards this goal, we present an empirical analysis of a corpus of breaking news stories that have been manually annotated for factual questions and answers. Our current goal is to compare extracted answers to a given question and to examine how features such as lexical similarity and source information relate to the chronological and semantic relationships between them. Our study will show that while there appears to be no direct relationship between the lexical similarity and publication time difference of a given answer pair, lexical similarity is highly correlated to whether or not the answers come from the same sources, and whether or not they express the same or different answer to the given question.

1 Introduction

When an important event happens, large numbers of news sources report on it. To do so, they draw information from direct participants in the event, eyewitnesses, official reports, as well as each other. As anyone who follows an event can attest, often multiple sources present complementary accounts of the news. Each source has its own techniques, informants, reputation, biases, and agenda. Sometimes a reader can get a fuller picture of an event only by consulting a number of complementary sources. News accounts of an event vary over time, in addition to source. Often initial reports turn out to be partially or fully incorrect. It takes time for accounts to stabilize and to be accepted

as facts. In such scenarios, where users wish to learn the correct answers to factual questions surrounding an important event, it is not appropriate simply to accumulate temporally disparate facts, or to use a voting scheme to ascertain the truth. Rather, one must explicitly incorporate the notions of source attribution and temporal extent, to account for change of both the ground truth as well as our (sources') knowledge of it.

1.1 Breaking news stories as evolving text

We conducted an initial analysis of three large clusters of emergency news stories, as reported by several Web-based agencies. The stories followed were the Columbia space shuttle disaster, the Rhode Island nightclub fire and the crash of a small plane into a skyscraper in Milan. (Attributes of the clusters can be found in Table 2.)

We read the most recently published article in each cluster, and generated a list of ten important facts central to each of the stories. We tracked the evolution of these factual questions across all documents in each cluster. We studied the relative order in which questions were answered and how long it took answers to stabilize (i.e. for all news sources to report the same information). In addition, we counted the number of times the answer to a question changed before stabilizing to the correct answer. This is shown in Table 1. It should be noted that 6 of the 30 questions never settled during the time period that the story made headlines. For example, in the RI fire story, two questions remained unresolved - who was to blame for the incident and whether or not the number of people inside the building at the time exceeded the legal capacity.

Among the 24 questions that did stabilize, the distribution of the time required to do so was rather skewed, with 8 questions taking longer than 24 hours, and 14 requiring less than 12 hours. For example, questions relating to the cause of an incident or the number of casualties are likely to stabilize over a longer period of time, while details external to the incident, such as the weather at the time of the event, are likely to settle relatively faster. In addition to the time to stabilization, another observation from our analysis is that certain facts in an evolving story are more volatile than others. For example, in the RI fire story, the answer to the question "How many victims were there?" changed 32 times before the correct answer was reported. The answer went from "at least 10," to "10 confirmed, actual feared much higher" to "several" to "at least 39" to "at least 60" and changed numerous times before reaching the final reported answer of "96 were killed."

Given the complexity of evolving news stories, we propose to track their key facts by automatically generating a matrix that summarizes the answers to a given factual question, over time and across a number of sources. This is illustrated in Figure 1 for one of the more volatile questions in the Milan plane crash story, "How many victims were there?" In the next section, we briefly discuss other areas of research that are related to our problem. In the remainder of the paper, we will present an empirical analysis of a larger corpus of emergency news clusters that have been manually annotated for answers to key factual questions. Specifically, our goal is to examine properties of the extracted answers that may change over time. In addition, we are interested in eventually predicting whether or not a set of extracted answers to a given question are essentially the same, or whether they are mutually exclusive, indicating that either

Order	Columbia shuttle breakdown			West Warwick, RI fire			Milan plane crash		
1	victims	1.5h	0	sprinklers	9.75h	0	height of building	3h	1
2	last contact	1.75h	0	fire code violation	12h	0	pilot killed	3.5h	0
3	terrorist act	1.75h	0	building description	15.5h	0	type of plane	3h	1
4	explosion	2h	4	injuries	24.75h	22	weather	4h	0
5	place	2h	2	cause	25h	9	passengers on plane	4h	1
6	location of debris	4h	6	fireworks permission	35.5h	14	plane's origin	8.5h	12
7	indications of trouble	14h	0	victims	35.5h	32	victims	24h	18
8	cause	57h	8	number in club	NA	NA	injuries	33h	13
9	parts found	59h	3	who was to blame	NA	NA	cause	NA	NA
10	injuries on ground	NA	NA	club over legal occupancy	NA	NA	number in building	NA	NA

Table 1: Relative order, time to stabilize and number of incorrect or partially correct answers before stabilization

the fact changed (in the physical world) or that there is disagreement between news sources. We introduce three specific hypotheses to be tested in Section 3.2.

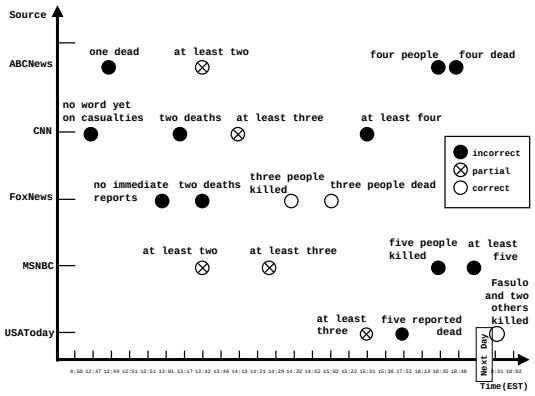

Figure 1: Timeline for the question "How many people were killed?" in the Milan story.

2 Related work

2.1 Changing information

The problem of tracking changing information over time is, of course, not new to the information retrieval community. For instance, the TDT First Story Detection task, in which the goal is to identify articles (in a stream of broadcast news) that introduce a new story, is an example of tracking information at the document level [2]. Similarly, other work has incorporated the notion of novelty detection into information filtering, with the goal of finding documents that not only fit a user's profile but that also contain novel information that the user has not yet seen [11].

Researchers have also looked at the problem of detecting new information at the

sentence level. One well-known research initiative is the TREC Novelty track, in which participants build systems that find the sentences (in a multi-document cluster of news) that are relevant to a given query and that contain "previously unseen information" [9]. However, as noted in both [9] and [4], one issue has been that when human judges are asked to annotate novel sentences, there is typically a large assessor effect (lower interjudge agreement).

We are proposing to track information at a finer level of granularity - the factual level. This approach is inspired by a previous study that found that judges were able to achieve very high levels of agreement (around 96%) in annotating facts in news stories [10]. Therefore, we are interested in studying how factual change is conveyed in text over time, which can inform our future work in building a "fact tracking system." This is somewhat similar in concept to δ-summarization, in which stories are tracked over time and summaries are produced in order to indicate to the user what has changed (since the time that the previous summary was produced) [1]. However, our system will track factual questions input by the user rather than try to find novel information in general.

2.2 Temporal relations in text

As noted by many working with clusters of related news articles, readers of a text must be able to determine when each event that is discussed happened in order to fully comprehend the text. However, events are not necessarily described in chronological order, particularly in narrative news stories [5]. Therefore, in order to develop a system for tracking changes in text over time, a system must be able to accurately resolve temporal relations in text.

Recently, Pustejovsky and colleagues noted that current question answering systems cannot support questions that refer to temporal aspects of events and entities in the world and their relative orderings [7]. An example of such a question is the following: "When did Iraq finally pull out of Kuwait during the war in the 1990s?" (p. 2). To this end, they developed TimeML, a metadata markup language for denoting event and temporal expressions in natural language corpora. Of interest are the four major data structures that are used in TimeML: events, time expressions (which might be explicit or relative), signals (such as prepositions that indicate how two objects are related) and links, which establish order between events. Unlike in [7], we will not attempt to markup all events in our corpus. Rather, we will focus on tracking specific facts about the events described in the stories, and will examine the properties that hold between the texts describing such facts.

3 Study setup

3.1 Corpus

We built a corpus of 9 emergency news stories, whose attributes are shown in Table 2. We included two types of document clusters in our corpus. "Newstrack" clusters were collected manually by the authors, who tracked a predefined set of online news outlets,

Story	Source	Documents	Questions	Answers	Sample question
Iraq suicide bombing	Newstrack	33	18	363	Who was the target of the attack?
Asian tsunami	Newstrack	146	5	40	Which countries were affected?
Milan plane crash	Newstrack	56	15	621	How many were injured?
RI nightclub fire	Newstrack	43	13	389	How many people were inside the building?
Columbia shuttle disaster	Newstrack	41	9	234	Where was debris found?
Gulfair plane crash	Newstrack	11	25	208	How many victims were there?
Kursk submarine disaster	Novelty-N33	25	20	211	Why did the Kursk sink?
Egyptair crash	Novelty-N4	25	22	265	Where did the plane crash?
China earthquake	Novelty-N43	25	8	106	What was the magnitude of the quake?

Table 2: Corpus of emergency news stories: cluster type, total number of documents, questions and extracted answers per cluster, and a sample question.

collecting all articles published about the story over a period of 48 hours. "Novelty" clusters were taken from the TREC Novelty track 2003 test data set [9].

For each cluster in the corpus, we asked one judge to read through the articles and to come up with a list of *factual* questions that are key to understanding that story. We collected between 15 and 30 questions for each story. Next, we assigned each cluster to another judge to find all answers to all of the questions. In particular, for each question, the judges went through every document in the assigned cluster, and found all answers to the question, listing the answer itself, as well as the document and sentence number where the answer was found [1]. In the instructions, the judges were told to find only explicit answers, but not those that give information that allow one to infer an answer. In some cases, judges found very few answers to a given question. Since we are interested in studying how answers change over time, we eliminated the questions with fewer than three answers from the data set. In total, our corpus consists of 135 factual questions annotated for answers (2,437 extracted answers in all). Once the answers were collected, one judge went through all sets of questions and answers and indicated, for each answer set to a given question, if the extracted answers expressed the same meaning or if the set contained some mutually exclusive answers.

3.2 Hypotheses

In the current paper, we will test three hypotheses that concern the relationship between vocabulary usage, and publication time and source.

H1: When comparing a pair of extracted answers to a given question, there is an inverse relationship between vocabulary overlap and publication time difference.

The first hypothesis to be tested concerns the relationship between vocabulary usage and publication time difference. We expect to see that in general, when answers are lexically similar to one another, the publication time difference between them (i.e. between their respective documents) is likely to be smaller as compared to answers that are lexically very dissimilar. This is because over longer periods of time, the fact of

[1] In a related study in which two judges found the sentences containing answers to questions in nine multi-document clusters, we reported a Kappa of 0.68 [6]

```
12:22 CNN: no word yet on casualities
12:42 MSNBC: no immediate report on casualties
14:29 MSNBC: at least three people killed
14:52 USA Today: killing at least three people
18:40 ABC News: leaving four dead
```

Figure 2: Examples of changing vocabulary over time for the question "How many victims were there?" in the Milan plane crash story.

interest is likely to have changed, resulting in the usage of new words. Figure 2 gives an example from the Milan plane crash cluster. It can be seen that the answers published within smaller time frames of one another are lexically more similar than those that have a large time difference between them.

H2: Answers to a given question that are extracted from different articles published by the same news source, have more shared vocabulary as compared to answers published by different sources.

Our second hypothesis concerns the relationship between the lexical similarity of extracted answers and whether or not they were published by the same news source. One reason why answers published by the same source might be more likely to be lexically similar to one another (as compared to those from different sources) is that journalists often use a system of rewrites when covering a breaking story, in which they simply update versions of previously published stories, adding only new information that has become available [5]. To contrast, given the widespread use of text from newswire services [3], we may find that there is not enough variation in vocabulary choice in order to distinguish between the answers published by different sources.

H3: Vocabulary overlap is higher in a set of extracted answers that are paraphrases of one another, versus a set in which there are mutually exclusive answers.

Finally, our third hypothesis considers the difference in vocabulary usage between sets of answers that express the same meaning versus those that contain mutually exclusive answers. While a set of answers with the same meaning may contain many paraphrases, we wish to test the hypothesis that on average, they exhibit a higher degree of lexical similarity than do a set containing mutually exclusive answers. By a set containing "mutually exclusive" answers, we mean a set of answers that could not be considered to report the same answer to the respective question. Figure 3 gives two examples of answer sets from our corpus that contain mutually exclusive answers. In the case of the Iraq suicide bombing example, the answers express different possible reasons for the attack. Similarly, in the Milan crash example, the answers contradict one another as to whether or not the crash was related to terrorism.

To contrast, Figure 4 shows some examples of answer sets that do not contain mutually exclusive answers. In the Iraq suicide bombing example, the answers refer to the same place in different ways. Similarly, in the first Egypt Air example, the answers

```
Iraq suicide bombing:
Q: What was the reason for the attack?
A1: to stop the party from participating
in the January election
A2: to intimidate the voters
A3: to threaten the voters
A4: to try to stop the election from
happening

Milan plane crash:
Q: Was it an accident?
A1: Marcello Pera said it "very probably"
appeared to be a terrorist attack.
A2: There were conflicting reports as to
whether it was a terrorist attack or an
accident.
A3: The crash appeared to be an accident.
A4: Authorities said it was an
apparent accident.
```

Figure 3: Examples of mutually exclusive answer sets.

refer to the same entity (the plane) differently. In the final Egypt Air example, the answers to the question are not mutually exclusive since one answers the question with an absolute temporal expression ("on Sunday") and the other does so with a related temporal expression ("20 minutes after..."). (As illustrated, the third example is one in which our hypothesis does not hold.)

3.3 Data sets

In order to test the three hypotheses, we created two data sets using our corpus of extracted answers. The first data set contains attributes for each of the 42,294 answer pairs (for a given question) that were compared. The second data set contains attributes of the 135 questions in the corpus and their respective answer sets.

First, for each question in the corpus, we compared the extracted answers pairwise with respect to five similarity metrics:

- **Simple cosine:** The cosine similarity using a binary count (1 if a word is shared between two answers, regardless of how many times, and 0 if not).
- **Cosine:** Cosine similarity using idf weights as well as the actual count of tokens in each extracted answer.
- **Token Overlap:** Proportion of shared tokens in both answers.
- **Norm. LCS:** Longest common substring normalized for answer length.

```
Iraq suicide bombing:
Q: Where did the attack take place?
A1: At the gate to the home of the leader
of Iraq's biggest political party.
A2: At the gate of Abdel-Aziz al-Hakim's
compound.
A3: At the gate at the home of Abdul
Aziz al-Hakim.

Egypt Air crash:
Q1: What kind of plane is the Boeing 767?
A1: Boeing 767-300ER
A2: a twin-engine jet
A3: a twin-engine, widebody passenger jet

Q2: When did the search mission begin?
A1: Sunday
A2: 20 minutes after the plane disappeared
from the radar screen
```

Figure 4: Three examples of answer sets that are not mutually exclusive.

In addition, we found the publication time difference (in minutes) between the answer pair, as well as whether or not they were published by the same news agency.

As potential control variables, we also included the expected answer type, as predicted by a manually created rule-based classifier used in our question answering system [8]. The expected answer types that appeared in our data set were the following: location, number, person, duration, reason, organization, biography, date distance, definition, place and other (those that did not fall into one of the other categories).

The second data set consists of attributes of each of the 135 questions in the corpus: the expected answer type, the total number of answers found by our judges for the question, the average pairwise similarity (for the five metrics mentioned above), average publication time difference and whether or not the set of extracted answers contains mutually exclusive answers.

4 Analysis

4.1 Hypothesis 1: lexical similarity and publication time difference

To test this hypothesis, we used the data set consisting of all pairwise comparisons of answers to questions in our corpus to fit a linear regression model with time difference as the response variable. The independent variables were the four similarity measures (simple cosine, cosine, token overlap and normalized LCS). In addition, we treated the following as control variables: the document cluster to which the answer pair con-

Variable	Corr. with TD
Cluster	-0.038
Answer type	-0.019
Same/diff source	0.021
Sim. cosine	0.038
Cosine	0.028
Token overlap	0.038
Norm. LCS	0.041

Table 3: Correlations between independent/control variables and publication time difference.

Indep. var.	P-value	Model R-square
Sim. cosine	0	0.0032
Cosine	0.00002	0.0032
Token overlap	0	0.0036
Norm. LCS	0	0.0037

Table 4: Regression of time difference on each similarity metric with cluster, source and answer type controlled.

cerned, the expected answer type, and whether or not the two answers were published by the same news source.

We first examined the correlations between the independent and control variables and the reponse variable, publication time difference. The correlation coefficients are shown in Table 3. Contrary to our expectations, all four of the similarity metrics have a slightly positive relationship with time difference.

Next, in order to examine the relationships between the similarity measures and time difference when the effects of source, cluster and answer type are controlled, we fit a regression model with each of the four metrics individually as the independent variable, along with the controls. We found that while all of the similarity metrics had a significant linear relationship to time difference, none of the models accounted for much of the variance in the response variable.

We also experimented with combining the independent variables and interactions between them or between them and the control variables. However, we did not find any model with an R-squared greater than 0.050, which would have little accuracy in predicting the time difference between a given pair of extracted answers. However, one interesting observation from the analysis is that the interaction terms between the source control variable (where 1 means the answers came from the same source and 0 indicates they came from difference sources) and all of the similarity metrics was always positive and significant .

We conclude that overall, there is a slight positive correlation between lexical similarity and time difference between answers to a given question, so we reject our original hypothesis. However, the source of the answers is an important confounding variable as is the expected answer type. In addition, we conclude that it is unlikely that we will be able to build a model to predict the publication time difference for a given pair of answers to a question, based on their lexical similarity, publishing source and expected answer type.

Similarity measure	Mean - same source	Mean - different sources	P-value
Simp. cosine	0.392	0.312	0
Cosine	0.392	0.312	0
Token overlap	0.327	0.232	0
Norm. LCS	0.355	0.264	0

Table 5: T-tests for the comparison of mean similarity between answer pairs published by the same news source vs. those published by different sources.

Attribute	Mean - not mut. exc.	Mean - mut. exc.	P-value
Answers found	13.8	22.8	0.005
Simp. cosine	0.578	0.334	0
Cosine	0.573	0.310	0
Token overlap	0.509	0.258	0
Norm. LCS	0.552	0.291	0

Table 6: T-tests for the comparison of mean similarity between answer pairs for questions in which there are not mutually exclusive answers vs. sets in which some answers are mutually exclusive.

4.2 Hypothesis 2: lexical similarity and news source

To test whether or not extracted answers published by the same news source are generally more lexically similar as compared to answer pairs from different sources, we conducted a t-test for each of the similarity metrics. The mean similarity between answers for each group (same source answers vs. those from different sources) and the p-value for the one-sided hypothesis test are shown in Table 5. Our conclusion with respect to our second hypothesis is that answer pairs published by the same source have more shared vocabulary than do answer pairs published by different news sources. This is true for all four of the metrics we tested.

4.3 Hypothesis 3: lexical similarity and mutual exclusivity of answer sets

To test our third hypothesis, we used the data set consisting of attributes of the 135 questions in the corpus. We divided the questions up into those that did not contain mutually exclusive answers and those that did. Our hypothesis is that answer sets containing mutually exclusive answers, on average, should exhibit less vocabulary overlap as compared to answer sets in which the same meaning is expressed. The average answer pair similarity, as well as the number of answers found per question, and the p-value for the t-test comparing the means between groups are shown in Table 6.

Clearly, on average, we can say that answers for a given question that are not mutually exclusive exhibit more lexical similarity as compared to answers from sets where some answers are mutually exclusive. In addition, the number of answers found for a question was typically greater in the sets containing mutually exclusive answers, as compared to the sets of answers expressing the same meaning.

5 Conclusion

In the current paper, we have presented an empirical analysis of a manually labeled corpus of factual questions and their corresponding answers in multi-document clusters of emergency news stories. Specifically, we tested three hypotheses that examined the relationship between the lexical similarity of the extracted answers and the chronological and source differences and semantic relationships between them. Our analysis suggests that there is no direct relationship between lexical similarity and publication time difference between a given pair of answers to a question, independent of other factors such as the source and the type of question. This is logical given that journalists often repeat information that has already been reported and the widespread use of newswire sources.

We did find clearer relationships between lexical similarity and source. On average in our corpus, answer pairs for a given question that are published by the same source are more similar than those coming from different sources. In addition, there was clearly more similarity between answer pairs that expressed the same meaning (were not mutually exclusive) as compared to those in which different meanings were expressed as an answer to the same question. In the future, we will focus on using a more semantic representation in comparing answers to a given question, rather than just lexical similarity. In particular, one direction for future work towards our goal of tracking facts over time, is to examine how discourse relationships between answer pairs to a question evolve with time [12] and whether or not there is again a large source effect.

References

[1] J. Allan, R. Gupta, and V. Khandelwal. Topic models for summarizing novelty. In *Proceedings of the Workshop on Language Modeling and Information Retrieval*, 2001.

[2] J. Allan, H. Jin, M. Rajman, C. Wayne, D. Gildea, V. Lavrenko, R. Hoberman, and D. Caputo. Topic-based novelty detection 1999 summer workshop at clsp final report, August 1999.

[3] P. Clough, R. Gaizauskas, S. S. Piao, and Y. Wilks. Measuring text reuse. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 152–159, 2002.

[4] D. Harman. Overview of the TREC 2002 novelty track, 2002.

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

[6] J. Otterbacher, G. Erkan, and D. R. Radev. Using random walks for question-focused sentence retrieval. In *Proceedings of HLT-EMNLP*, Vancouver, BC, 2005.

[7] J. Pustejovsky, J. Castano, R. Ingria, R. Saurí, R. Gaizauskas, A. Setzer, G. Katz, and D. Radev.

[8] D. R. Radev, W. Fan, H. Qi, H. Wu, and A. Grewal. Probabilistic Question Answering from the Web. In *The 11th International World Wide Web Conference*, Honolulu, Hawaii, May 2002.

[9] I. Soboroff and D. Harman. Overview of the TREC 2003 Novelty Track. In *Proceedings of the Twelfth Text Retrieval Conference (TREC 2003)*, NIST, Gaithersburg, ML, 2003.

[10] H. van Halteren and S. Teufel. Examining the Consensus Between Human Summaries: Initial Experiments with Factoid Analysis. In *Proceedings of HLT-NAACL 2003 Workshop on Text Summarization (DUC03)*, Edmonton, 2003.

[11] Y. Zhang, J. Callan, and T. Minka. Novelty and Redundancy Detection in Adaptive Filtering. In *Proceedings of SIGIR 2002*, 2002.

[12] Z. Zhang and D. R. Radev. Learning cross-document structural relationships using both labeled and unlabeled data. In *Proceedings of IJC-NLP 2004*, Hainan Island, China, March 2004.