Detecting Discrepancies in Numeric Estimates Using Multidocument Hypertext Summaries

Michael White
CoGenTex, Inc.
840 Hanshaw Road
Ithaca, NY 14850, USA
mike@cogentex.com

Claire Cardie
Dept. of Computer Science
Cornell University
Ithaca, NY 14850, USA
cardie@cs.cornell.edu

Vincent Ng
Dept. of Computer Science
Cornell University
Ithaca, NY 14850, USA
yung@cs.cornell.edu

ABSTRACT

To aid analysts in detecting discrepancies in numeric estimates in news articles from multiple sources, we propose the automatic generation of hypertext summaries that include a high-level textual overview; tables of all *comparable* numeric estimates, organized to highlight discrepancies; and targeted access to supporting information from the original articles. The RIPTIDES system, which exemplifies the more flexible human-computer interface we propose, combines information extraction and multidocument summarization techniques to produce such hypertext summaries. In evaluating the system's ability to facilitate discrepancy detection, we find that, on average, the hypertext summaries provide a significantly more complete picture of the available information than the latest article.

1. INTRODUCTION

Previous work in multidocument summarization has pointed to the importance of identifying differences and discrepancies in the information that is reported across multiple news sources [9, 12]. To our knowledge, however, this problem has not yet been systematically or thoroughly investigated. Radev and McKeown [9], for example, identify discrepancy detection as a potential problem for multidocument summarizers via anecdotal evidence, but provide no empirical evidence to indicate how often such differences actually represent significant discrepancies in the available information, vs. simple updates in what is known. In particular, it is unclear whether readers may usually find a complete and accurate picture of the available information by simply looking at the latest article.

Similarly, our recent case study in the domain of natural disasters as part of the DUC summarization evaluation only begins to investigate issues of discrepancy detection [12]. The study attempted to (1) quantify the need for detecting discrepancies in numeric estimates of injury and death tolls, and (2) evaluate the ability of automatically generated summaries to deal with discrepancies and provide a more complete and accurate picture of an ongoing event than is available in the latest article. The study confirmed our impression that although the estimates do usually converge, they change rapidly at first and are often dropped from later articles, making the latest article an unreliable source for this information. We also found that manually scanning articles from multiple news sources for the latest death and injury estimates is exceedingly tedious. Further details from this case study are presented in the next section.

.

While our case study indicated that discrepancy detection, at least in the domain of natural disasters, is a task that could certainly benefit from automation, we unfortunately also found that the previous version of our RIPTIDES summarizer [12, 13] could not help identify such discrepancies more reliably than the latest article, on average. The conclusion we drew from this study was that length-limited, generic textual summaries necessarily preclude the inclusion of a complete and detailed assessment of discrepancies, while simultaneously making it difficult for the end-user to identify any inaccurate estimates in the summary.

We hypothesize here, therefore, that complete, accurate, and easily absorbed numeric estimates would be better conveyed to analysts in multidocument summaries that espoused a more flexible human-computer interface. For this, we propose the automatic generation of hypertext summaries that include a high-level textual overview; tables of all *comparable* numeric estimates, organized to highlight discrepancies; and targeted access to supporting information from the original articles.

In the sections below, we present and evaluate the new and improved RIPTIDES system and its hypertext summarization capability, which exemplifies the proposed more flexible human-computer interface. Our evaluation shows that, on average, the hypertext summaries provide a significantly more complete picture of the available information than the latest article. Most strikingly, when compared to a subset of 10 articles with incomplete reporting of the available information on the death toll, the RIPTIDES summarizer scores 4.55 on a 5-point scale in its completeness of death toll reporting, vs. only 2.05 for the selected articles.

2. CASE STUDY

In contrast to many available document collections that have been used in multidocument summarization research (e.g. the DUC and TDT corpora [7]), which consist of relatively small sets of articles about a given topic, our study instead required a fairly complete set of news articles about a single, evolving event. As a result, we created a corpus of as many articles from the web as we could easily find during the first week after the January 2001 earthquake in Central America: 132 articles from five news sources — AP, Reuters, CNN, BBC and the Washington Post. We then examined articles from days one through four of the quake to see how often the most recent article gave a signficantly different picture of the death toll than one would obtain from reading all the articles up to that point. We found that 20% (22/107) of the articles failed to provide an accurate picture of what was known about the death toll at the time, containing either a significant discrepancy or no information on the overall death toll whatsoever (half of the 22 articles). For example, a BBC article on the second day reported a death toll of at least 80, which was consistent with the latest confirmed estimates from CNN and Reuters, but conflicted with another BBC article (posted one minute earlier) that gave an estimate of hundreds, as well as with the latest AP estimate of at least 122, and a quote from a police agency in the same AP article of 234.

We also examined the leading two paragraphs of all articles, and found significant variation in the facts reported, suggesting that there is considerable opportunity for a multidocument summarizer to surface key facts that may be missing from the most recent article leads. For example, while the death toll was mentioned in the first two paragraphs in 72% of the articles (95/132), the number missing appeared in only 41% (54/132), and the number injured in only 10% (13/132). In contrast, upon examining the full text of the articles, we found that the death toll was mentioned in about 92% of the articles, and both the number missing and injured in around 60% of the articles.

3. SYSTEM DESCRIPTION

RIPTIDES combines information extraction (IE) and multidocument summarization techniques to produce its domain-specific hypertext summaries. First, the system requires (1) the selection of one or more scenario templates (extraction domains), and (2) a set of documents in which to search for information. Our current study uses the domain of natural disasters and the Central American Quake (CAQ) corpus described above, respectively. RIPTIDES then applies its IE subsystem to generate a database of extracted events for the selected domain and invokes the Summarizer to generate a hypertext summary of the extracted information. The next subsection walks through sample hypertext summaries, pointing out the key features of the RIPTIDES interface. It is followed by descriptions of the IE and Summarizer system components.

3.1 Examples

The hypertext summaries consist of a high-level textual overview plus an indexed set of tables of all extracted information. Figure 1 shows a textual overview of the first dozen or so articles in the CAQ corpus. The 200-word overview contains sentences extracted from the original articles. The selection of sentences to extract is designed to favor adjacent sentences, in order to improve the intelligibility of the resulting summary; in figure 1, three blocks of adjacent sentences are shown. Clicking on the magnifying glass icon brings up the original article in the right frame, with the extracted sentences highlighted.

The index to the hypertext summary appears in the left frame of figure 1. Links to the overview and lead sentence of each article are followed by links to summary information organized according to the base level extraction slots for the main event (here, an earthquake) including its description, date, location, epicenter and magnitude. Access to overall damage estimates appear next, with separate tables for types of human effects (e.g. dead, missing) and for object types (e.g. villages, bridges, houses) with physical effects.

Figure 2 shows the extracted estimates of the overall death toll. In order to help identify discrepancies, the high and low current estimates are shown at the top, followed by other current estimates and then all extracted estimates. Heuristics are used to determine which estimates to consider current, taking into account the source (either news source or attributed source), specificity (e.g. *hundreds* vs. *at least 200*) and confidence level, as indicated by the presence of hedge words such as *perhaps* or *assumed*. The tables also provide links to the original articles, allowing the user to quickly and directly determine the accuracy of any estimate in the table. By following the high-low estimate links in figure 2, for example, one would discover that the lower estimates (of at least two dead) appear to only take the Guatemala estimates into account, whereas the higher estimates also include the reported death toll from Las Colinas, near San Salvador.

Figure 3 shows a more substantial index of tables. The input articles for this example are from topic 89 of the TDT2 corpus, a set of newswires that describe the May 1998 earthquake in Afghanistan. As above, the index begins with the overview, the article leads, entries for each extraction slot in the main event, and overall damage estimates. Here, however, the overall estimates are followed by estimates for specific locations, such as Shari Basurkh or Rustaq. The same indexing pattern is then followed for any related events mentioned in the articles, such as aftershocks, landslides, or previous quakes in the same area.

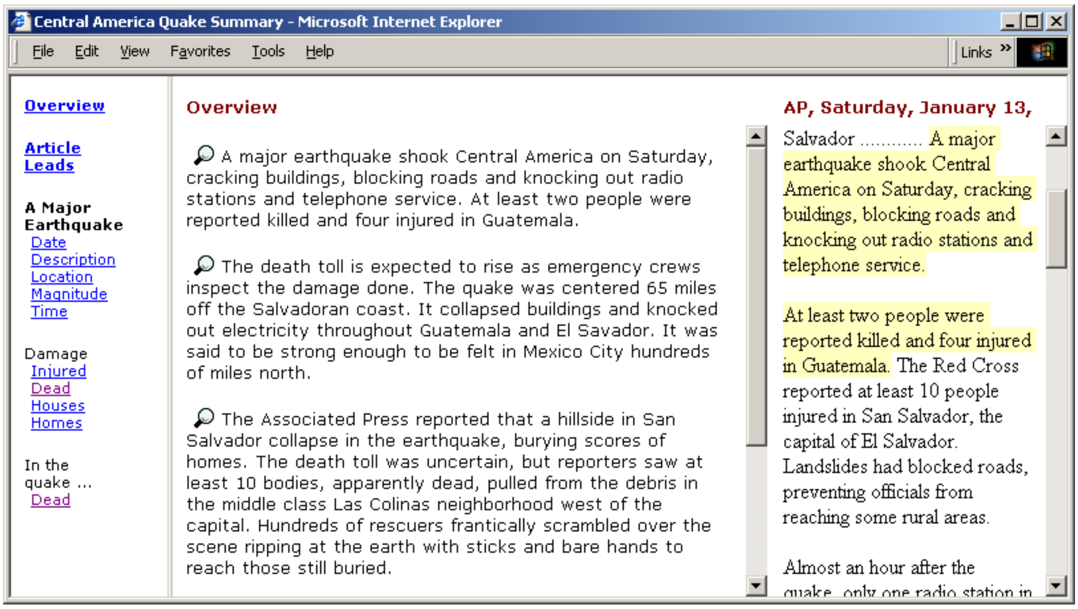
3.2 IE System

The RIPTIDES IE system combines existing language technology components in a traditional system architecture [2]: named entity identification via BBN's Identifinder [1]; sentence boundary detection and date normalization via Mitre's Alembic Workbench [4]; syntactic parsing via the Charniak [3] parser; WordNet-based [5] selectional restrictions; and linguistic annotation management via an in-house implementation of the TIPSTER architecture [6]. Unique features of the system include a weakly supervised extraction pattern-learning component, Autoslog-XML, which is based on Autoslog-TS [10], but operates in an XML framework and acquires patterns for extracting text elements beyond noun phrases, e.g. verb groups, adjectives, adverbs, and single-noun modifiers. In addition, a heuristic-based clustering algorithm organizes the extracted concepts into output templates specifically designed to support multi-document summarization [13]: the IE system, for example, distinguishes different reports or views of the same event from multiple sources.

Figure 4 shows the output (on the left) of the information extraction system given a sample input text from topic 89 of the TDT2 corpus (on the right). Output templates from the IE system for each text to be covered in the multi-document summary are provided as input to the summarization component along with all linguistic annotations accrued in the IE phase.

3.3 Summarizer

The Summarizer operates in three main stages. In the first stage, the IE output templates are merged into an event-oriented structure where comparable facts are semantically grouped. Towards the same objective, surface-oriented clustering is used to group sentences from different documents into clusters that are likely to report similar content. In the second stage, importance scores are assigned to the sentences based on the following indicators: position in document, document recency, presence of quotes, average sentence overlap, headline overlap, size of cluster (if any), size of semantic groups (if any), specificity of numeric estimates, and whether these estimates are deemed current. In the third and final stage, the hypertext summary is generated from the resulting content pool. A stochastic search procedure is used to select the highest ranking set of sentences for the overview summary; in this search, the inclusion of adjacent sentences is favored and the selection of repetitive material is penalized, in order to improve intelligibility. Further details on each stage follow in the paragraphs below.



Overview

Article Leads

A Major Earthquake
Date
Description
Location
Magnitude
Time

Damage
Injured
Dead
Houses
Homes

In the quake ...
Dead

Overview

🔍 A major earthquake shook Central America on Saturday, cracking buildings, blocking roads and knocking out radio stations and telephone service. At least two people were reported killed and four injured in Guatemala.

🔍 The death toll is expected to rise as emergency crews inspect the damage done. The quake was centered 65 miles off the Salvadoran coast. It collapsed buildings and knocked out electricity throughout Guatemala and El Savador. It was said to be strong enough to be felt in Mexico City hundreds of miles north.

🔍 The Associated Press reported that a hillside in San Salvador collapse in the earthquake, burying scores of homes. The death toll was uncertain, but reporters saw at least 10 bodies, apparently dead, pulled from the debris in the middle class Las Colinas neighborhood west of the capital. Hundreds of rescuers frantically scrambled over the scene ripping at the earth with sticks and bare hands to reach those still buried.

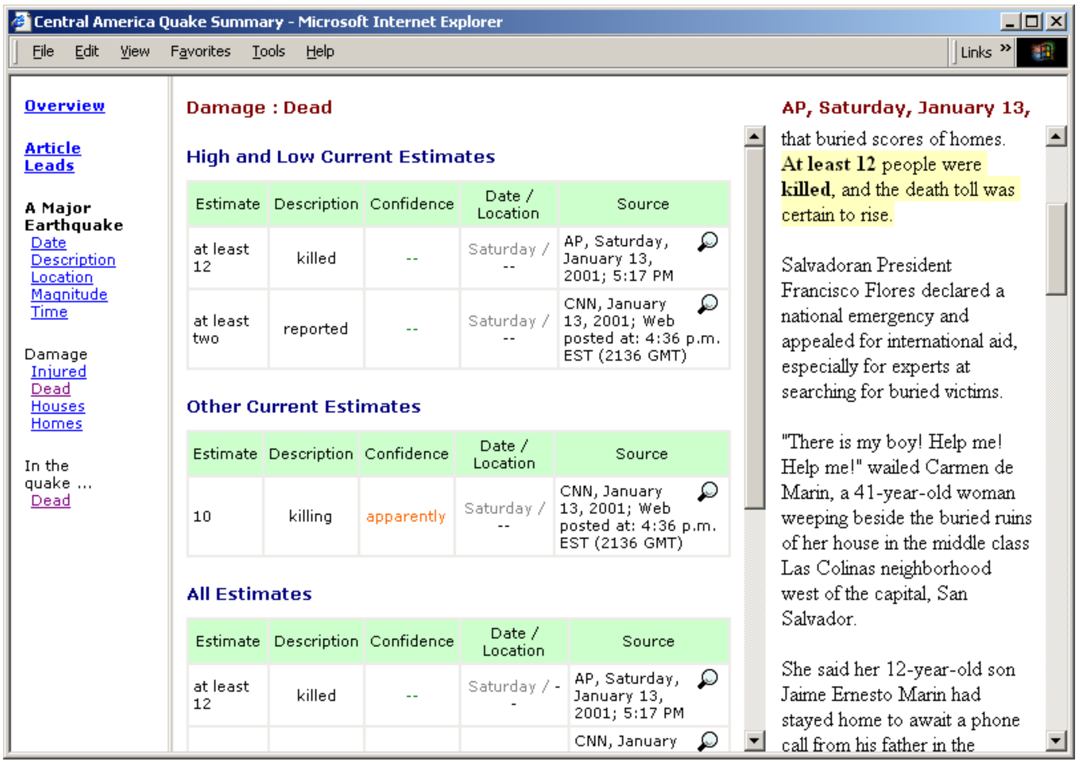
AP, Saturday, January 13,

Salvador .... A major earthquake shook Central America on Saturday, cracking buildings, blocking roads and knocking out radio stations and telephone service.

At least two people were reported killed and four injured in Guatemala. The Red Cross reported at least 10 people injured in San Salvador, the capital of El Salvador. Landslides had blocked roads, preventing officials from reaching some rural areas.

Almost an hour after the quake, only one radio station in

Figure 1: Example Multidocument Hypertext Summary Overview



Overview

Article Leads

A Major Earthquake
Date
Description
Location
Magnitude
Time

Damage
Injured
Dead
Houses
Homes

In the quake ...
Dead

Damage : Dead

High and Low Current Estimates

Estimate	Description	Confidence	Date / Location	Source
at least 12	killed	--	Saturday / --	AP, Saturday, January 13, 2001; 5:17 PM 🔍
at least two	reported	--	Saturday / --	CNN, January 13, 2001; Web posted at: 4:36 p.m. EST (2136 GMT) 🔍

Other Current Estimates

Estimate	Description	Confidence	Date / Location	Source
10	killing	apparently	Saturday / --	CNN, January 13, 2001; Web posted at: 4:36 p.m. EST (2136 GMT) 🔍

All Estimates

Estimate	Description	Confidence	Date / Location	Source
at least 12	killed	--	Saturday / --	AP, Saturday, January 13, 2001; 5:17 PM 🔍
				CNN, January 🔍

AP, Saturday, January 13,

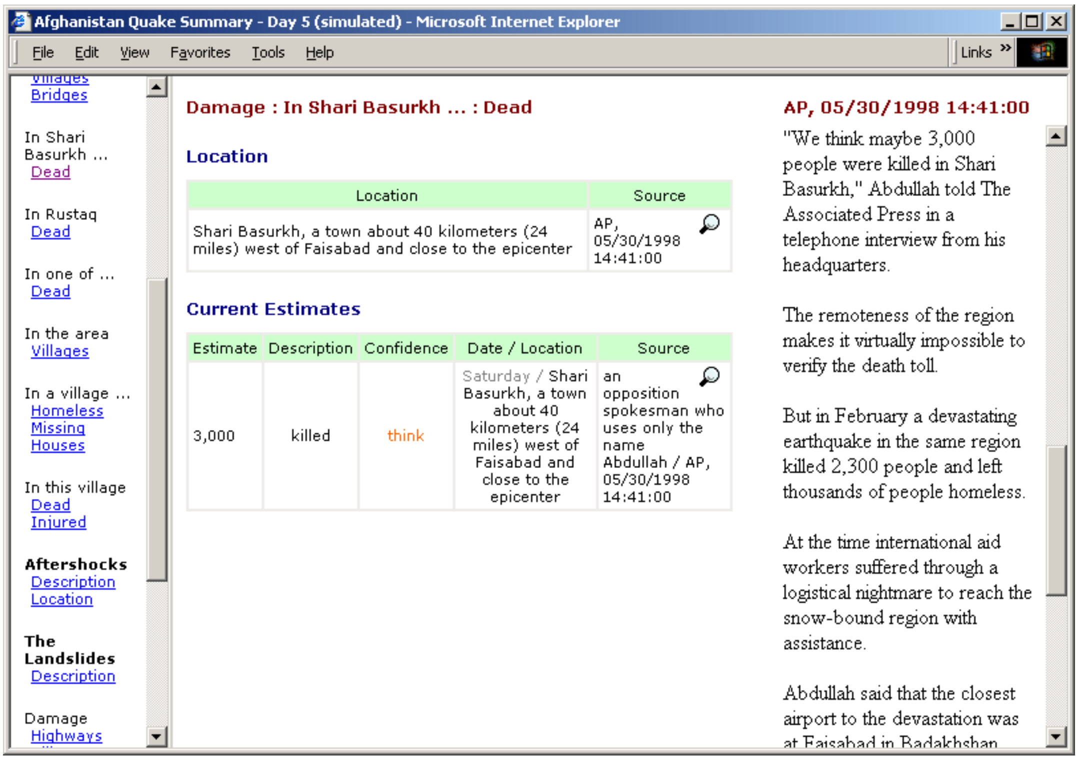
that buried scores of homes. **At least 12** people were **killed**, and the death toll was certain to rise.

Salvadoran President Francisco Flores declared a national emergency and appealed for international aid, especially for experts at searching for buried victims.

"There is my boy! Help me! Help me!" wailed Carmen de Marin, a 41-year-old woman weeping beside the buried ruins of her house in the middle class Las Colinas neighborhood west of the capital, San Salvador.

She said her 12-year-old son Jaime Ernesto Marin had stayed home to await a phone call from his father in the

Figure 2: Example Tables of Death Toll Estimates


Afghanistan Quake Summary - Day 5 (simulated) - Microsoft Internet Explorer
File
Edit
View
Favorites
Tools
Help
Links

Villages
Bridges

In Shari Basurkh ...
Dead

In Rustaq
Dead

In one of ...
Dead

In the area
Villages

In a village ...
Homeless
Missing
Houses

In this village
Dead
Injured

Aftershocks
Description
Location

The Landslides
Description

Damage
Highways

Damage : In Shari Basurkh ... : Dead

Location

Location	Source
Shari Basurkh, a town about 40 kilometers (24 miles) west of Faisabad and close to the epicenter	AP, 05/30/1998 14:41:00

Current Estimates

Estimate	Description	Confidence	Date / Location	Source
3,000	killed	think	Saturday / Shari Basurkh, a town about 40 kilometers (24 miles) west of Faisabad and close to the epicenter	an opposition spokesman who uses only the name Abdullah / AP, 05/30/1998 14:41:00

AP, 05/30/1998 14:41:00

"We think maybe 3,000 people were killed in Shari Basurkh," Abdullah told The Associated Press in a telephone interview from his headquarters.

The remoteness of the region makes it virtually impossible to verify the death toll.

But in February a devastating earthquake in the same region killed 2,300 people and left thousands of people homeless.

At the time international aid workers suffered through a logistical nightmare to reach the snow-bound region with assistance.

Abdullah said that the closest airport to the devastation was at Faisabad in Badakhshan

Figure 3: Example Index With Groupings by Type, Location and Related Event

Document no.: ABC19980530.1830.0342
Date/time: 05/30/1998 18:35:42.49

Disaster Type: earthquake
- description: *a powerful earthquake*
- location: *Afghanistan*
- date: *today*
- magnitude: *6.9*
- magnitude-confidence: high
- epicenter: *a remote part of the country*
- damage:
  - human-effect:
    - victim: *Thousands of people*
    - number: *Thousands*
    - outcome: dead
    - description: *dead*
    - confidence-marker: *feared*
    - confidence: medium
  - physical-effect:
    - object: *entire villages*
    - outcome: damaged
    - description: *buried*
    - confidence: medium
    - confidence-marker: *Details now hard to come by / reports say*

PAKISTAN MAY BE PREPARING FOR ANOTHER TEST

Thousands of people are feared dead following... (voice-over) ...a powerful earthquake that hit Afghanistan today. The quake registered 6.9 on the Richter scale, centered in a remote part of the country. (on camera) Details now hard to come by, but reports say entire villages were buried by the quake.

Figure 4: Information Extraction in the Domain of Natural Disasters. Output of the system is shown on the left; input, on the right.

In the analysis stage, we use Columbia's SimFinder tool [8] to obtain surface-oriented similarity measures and clusters for the sentences in the input articles. To obtain potentially more accurate partitions using the IE output, we semantically merge the extracted slots into *comparable* groups, i.e. ones whose members can be examined for discrepancies. This requires distinguishing (i) different types of damage, according to outcome (e.g. injured, dead, missing) for human effects, and according to object type (e.g. houses, vehicles) for physical effects; (ii) overall damage estimates vs. those that pertain to a specific locale; and (iii) damage due to related events, such as previous quakes in the same area. For example, we want to separate out overall damage estimates for the main event from ones that pertain to a specific locale or to a previous quake, as these estimates cannot be directly compared to see if they conflict.

The merging routine relies on simple domain- and genre-specific heuristics to group slots across the input documents. Devising a more accurate and robust merging routine remains an interesting topic for future research. The current routine assumes there is a single main event which all the documents are about, and that there is no more than one sub-event or related event for each event type (i.e. landslide, aftershock, earthquake, etc.). To determine when a damage estimate is specific to a certain locale, we ran a decision tree induction algorithm on the manually labeled damage estimates from TDT2 topic 89, and were surprised to find that sentence position is a surprisingly strong predictor; indeed, on this data set, classifying an estimate as localized whenever it had a location and appeared in the fifth sentence or beyond turned out to be 99% correct on the training data.

During the analysis stage, we also analyze the numeric estimates for specificity, using a small set of patterns over POS tags. Three levels of specificity are identified: rough estimates (e.g. *hundreds*), ranges and exact figures. Some numeric estimates, such as percentages, remain unanalyzed. The patterns identify the magnitude of the rough estimates and the lower and upper bounds of the ranges, so that the high and low estimates in a group can be identified. After the numeric estimates have been analyzed and confidence levels assigned, the current estimates are identified. In determining when to consider an estimate current, a later report from the same source (news agency or attributed source) is assumed to supercede an earlier one when it is at least as specific or higher, and has at least the same confidence level.

In the scoring stage, SimFinder's similarity measures and clusters are combined with the semantic groupings obtained from merging the IE templates in order to score the input sentences. The scoring of the clusters and semantic groups is based on their size, and the scores are combined at the sentence level by including the score of all semantic groups that contain a phrase extracted from a given sentence. More precisely, the scores are assigned in two phases, according to a set of hand-tuned parameter weights. First, a base score is assigned to each sentence according to a weighted sum of the position in document, document recency, presence of quotes, average sentence overlap, and headline overlap. The average sentence overlap is the average of all pairwise sentence similarity measures; we have found this measure to be a useful counterpart to sentence position in reliably identifying salient sentences, with the other factors playing a lesser role. In the second scoring phase, the clusters and semantic groups are assigned a score according to the sum of the base sentence scores. After normalization, the weighted cluster and group scores are used to boost the base scores, thereby favoring sentences from the more important clusters and semantic groups. Finally, a small boost is applied for current and more specific numeric estimates.

In the generation stage, the overview is constructed by selecting a set of sentences in a context-sensitive fashion, then ordering the blocks of adjacent sentences according to their importance scores. The scoring model begins with the sum of the scores for the candidate sentences, which is then adjusted to penalize the inclusion of multiple sentences from the same cluster or semantic group, or sentences whose similarity measure is above a certain threshold, and to favor the inclusion of adjacent sentences from the same article, in order to boost intelligibility. A larger bonus is applied when including a sentence that begins with an initial pronoun as well as the previous one, and an even bigger bonus is added when including a sentence that begins with a strong rhetorical marker (e.g. *however*) as well as its predecessor; corresponding penalties are also used when the preceding sentence is missing, or when a short sentence appears without an adjacent one.

To select the sentences for the overview according to this scoring model, we use a simple stochastic search method, namely randomized local search from multiple starting points (cf. [11]). For the first iteration, we begin with the highest scoring sentences up to the word limit. For subsequent iterations, we begin with randomly selected sentences, weighted according to their scores, up to the word limit. During each iteration, a random step or a greedy step is repeatedly performed until a greedy step fails to improve upon the current set of sentences. In each random step, a randomly selected sentence is added to collection. In each greedy step, one sentence is chosen to add to the summary, and zero or more (typically one) sentences are chosen to remove from the summary, such that the word limit is still met, and this combination of sentences represents the best swap available according to the scoring model. The search continues for a predetermined number of iterations, keeping track of the best combination of sentences found so far; we have found that 10 iterations often suffices to find a reasonable collection.

Once the overview sentences have been selected, the hypertext summary is generated as a collection of HTML files, using a series of XSLT transformations.

3.4 Training and Tuning

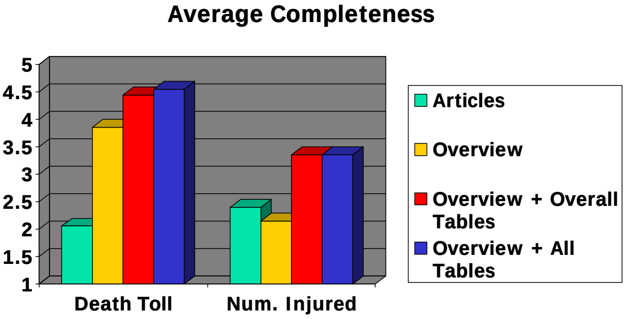
For the evaluation below, the IE system was trained on 12 of 25 texts from topic 89 of the TDT2 corpus. It achieves 42% recall and 61% precision when evaluated on the remaining 13 topic 89 texts. The parameters of the Summarizer were chosen by hand using the TDT2 topic 89 document set.

4. EVALUATION METHOD AND RESULTS

To determine the inputs for the evaluation, we selected 10 of the first 22 articles in the CAQ corpus that failed to completely and accurately report the available information on the overall death toll, since these represent some of the more interesting cases where automatic support for detecting discrepancies would be most useful. For each article, we then ran the RIPTIDES system on the articles up to and including that article (with a limit of 20), producing overview summaries of 200 words or less, as well as a series of hyperlinked tables of damage estimates. Next we had two judges[1] rate each selected article and its corresponding hypertext summary on the completeness of its reporting of both the overall death toll and the overall number injured. The ratings were given on a five point scale, where 1 = 'not at all,' 2 = 'somewhat,' 3 = 'fairly,' 4 = 'mostly,' and 5 = 'entirely.'

Table 5 shows the results averaged across the two judges. The 200-word overviews scored significantly better than the selected articles in their death toll reporting, with an average completeness score of 3.85 vs. 2.05 ($p < 0.001$, two-tailed paired t-test). The

[1] The first two authors were the judges.



	Death Toll	Num. Injured
Articles	2.05	2.40
Overview	3.85	2.15
Overview + Overall Tables	4.45	3.35
Overview + All Tables	4.55	3.35

Figure 5: Results of Discrepancy Detection Evaluation

overviews scored slightly worse on their reporting of the number injured (2.15 vs. 2.40), but this difference was not significant. When we looked at the overviews in combination with the tables of extracted estimates for the overall death toll and the overall number injured, the system fared even better, scoring 4.45 (vs. 2.05) for the death toll, and 3.35 (vs. 2.40) for the number injured ($p = 0.011$). Looking further to all the death toll tables raised the system's score another notch to 4.55, as these tables contained a few current estimates that were mistakenly grouped with reports for specific locations or previous disasters.

We consider these results to be quite promising, as they show that the system is able to identify discrepancies in numeric estimates reasonably well even on hard cases. While the information analyst must examine the hypertext tables carefully in order to separate out the best available estimates from those that have been superceded or are incorrectly classified, we have found that it is still much easier to scan the tables for this information than to scan the full text of the original articles.

5. ACKNOWLEDGMENTS

This work was supported in part by DARPA TIDES contract N66001-00-C-8009.

6. ADDITIONAL AUTHORS

Daryl McCullough, CoGenTex, Inc., daryl@cogentex.com.

7. REFERENCES

[1] D. Bikel, S. Miller, R. Schwartz, and R. Weischedel. Nymble: A High-Performance Learning Name-Finder. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pages 194–201, San Francisco, CA, 1997. Morgan Kaufmann.

[2] C. Cardie. Empirical Methods in Information Extraction. *AI Magazine*, 18(4):65–79, 1997.

[3] E. Charniak. A maximum-entropy-inspired parser. Technical Report CS99-12, Brown University, 1999.

[4] D. Day, J. Aberdeen, L. Hirschman, R. Kozierok, P. Robinson, and M. Vilain. Mixed-Initiative Development of Language Processing Systems. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*. Association for Computational Linguistics, 1997.

[5] C. Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, 1998.

[6] R. Grishman. TIPSTER Architecture Design Document Version 2.2. Technical report, DARPA, 1996. Available at `http://www.tipster.org/`.

[7] D. Harman. *Proceedings of the 2001 Document Understanding Conference (DUC-2001)*. NIST, 2001.

[8] V. Hatzivassiloglou, J. L. Klavans, M. L. Holcombe, R. Barzilay, M.-Y. Kan, and K. R. McKeown. Simfinder: A flexible clustering tool for summarization. In *Proceedings of the NAACL 2001 Workshop on Automatic Summarization*, Pittsburgh, PA, 2001.

[9] D. R. Radev and K. R. McKeown. Generating natural language summaries from multiple on-line sources. *Computational Linguistics*, 24(3):469–500, 1988.

[10] E. Riloff. Automatically Generating Extraction Patterns from Untagged Text. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, pages 1044–1049, Portland, OR, 1996. AAAI Press / MIT Press.

[11] B. Selman and H. Kautz. Noise Strategies for Improving Local Search. In *Proceedings of AAAI-94*, 1994.

[12] M. White, C. Cardie, V. Ng, K. Wagstaff, and D. McCullough. Detecting discrepancies and improving intelligibility: Two preliminary evaluations of RIPTIDES. In *Proceedings of the 2001 Document Understanding Conference (DUC-2001)*, 2001.

[13] M. White, T. Korelsky, C. Cardie, V. Ng, D. Pierce, and K. Wagstaff. Multidocument Summarization via Information Extraction. In *Proceedings of the First International Conference on Human Language Technology Research*, San Diego, CA, 2001.