Machine and Human Performance for Single and Multidocument Summarization

Judith D. Schlesinger and John M. Conroy, *IDA/Center for Computing Sciences*

Mary Ellen Okurowski, *Department of Defense*

Dianne P. O'Leary, *University of Maryland*

The DUC 2002 evaluation revealed numerous language-processing challenges that impact text summarization. This article examines the techniques used in a multidocument summarization system the authors developed and its performance at DUC 2002. It also discusses the need for regularization of human summaries.

Automatic multidocument summarization poses interesting challenges to the natural language processing (NLP) community. Multidocument summary generating systems must both cover single-document summarization issues—such as determining the generated summary's relevant information, pronoun resolution, and coherency—and be able to draw the "best" information from a set of documents.

Automatic single-document text summarization[1] has been an active research area since the 1950s, with a renaissance of approaches since the 1990s. Human single-document summarization is well defined when guidelines and recommendations drive performance.[2,3] System-generated single-document summaries, while not always matching well with reference summaries, are generally good quality and have proved useful. In contrast to the single-document task, multiple-document summarization development (automated or not) lacks complementary documentation of procedures and methodologies for human performance. Although researchers have explored various strategies for analyzing documents in a collection and then synthesizing and condensing information to produce multidocument summaries, they have not yet seen strong system performance. The lack of guidelines has a greater impact on multidocument summarization than on the single-document task. Analyzing and synthesizing such extensive information tax both human and machine-processing capabilities.

The recent NIST-sponsored (National Institute of Standards and Technology) Document Understanding Conference II (DUC 2002)[4] evaluated automatic multidocument summarization capability using 59 single-document collections with an average of 10 documents per set. These sets covered single events, multiple related events, and biographies. NIST provided 30 document sets, including single documents, single-document human-generated reference summaries, and multidocument human-generated reference summaries for each set. Additional multidocument summaries for each set, which different human summarizers wrote, were also available, as was data from the DUC 2001 conference. For evaluation, human evaluators compared system-generated summaries to the reference human summaries.

Our prototype multidocument summarizer operates by first generating single-document summaries and then selecting from those sentences to produce the multidocument summary. It is based on our current text summarization system, which delivers, in real time, indicative summaries for a high-volume, heterogeneous document collection to US government users. Thus, the constraints of our environment tend to prohibit the use of more knowledge-intense approaches but truly represent typical requirements for commercial viability. We believe our environment highlights actual practical performance demands and NLP challenges.

Generating single-document abstracts

Our algorithm for generating single-document abstracts is based on a hidden Markov model (HMM).

We trained the model using a mapping of NIST-

1094-7167/03/$17.00 © 2003 IEEE
Published by the IEEE Computer Society

provided summaries to their source sentences, and we evaluated our approach by analyzing the relation of human multidocument summaries to single-document ones and by assessing coverage of the single-document summaries.

The algorithm

In contrast to a naive Bayesian approach,[5,6] an HMM has fewer assumptions of independence (for more about HMMs, see the "Further Reading" sidebar on page 52). Particularly, it does not assume that the probability that sentence i is in the summary is independent of whether sentence $i - 1$ is in the summary. We also use a joint distribution for the feature set.

Our feature set is quite shallow, combining surface-level and entity-level approaches (for a description of these approaches, see the "NLP Features for Summarization" sidebar). Our tokenizer employs the robust SRA NetOwl software with named-entity recognition and aliasing. The set includes

- The sentence's position in the document—built into the HMM's state structure
- The number of tokens (non-stop words) in the sentence—value is $o_1(i) = \log(\textit{number\_of\_tokens} + 1)$
- The number of *pseudo-query* terms in a sentence—$o_2(i) = \log(\Pr(\log(\textit{number\_of\_pseudo-query\_terms} + 1)))$

Pseudo-query terms, also called *signature terms*, are terms that are more likely to occur in the document (or document set) than in the corpus at large. For query-driven summaries, the users define these terms and phrases (for more on query-driven summaries, see the "Further Reading" sidebar). We omit an additional feature—a sentence's distance from one that contains at least one query term—when generating generic summaries because nearly every sentence contains at least one pseudo-query term. To identify these terms, we use the log-likelihood statistic suggested by Ted Dunning,[7] which is equivalent to a mutual information statistic. The statistic is based on a two-by-two contingency table of counts for each term.

We expected the probability that the next sentence was included in the summary to change depending on whether the current sentence was a summary sentence. A first-order Markov model allowed such differences with marginal additional cost over a simple Bayesian classifier.

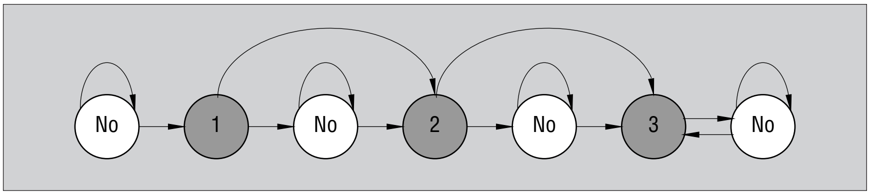

Figure 1. Summary extraction Markov model for extracting two lead sentences and additional supporting sentences.

An HMM handled the positional dependence, feature dependence, and Markovity. Our model has $2s + 1$ states, with s summary states and $s + 1$ nonsummary states. Figure 1 shows the Markov chain.

In the Markov model, sentences processed by a numbered state become part of the summary, while those processed by a state labeled "no" do not. We allow hesitation only in nonsummary states and the skipping of states only in summary states. We designed this chain to model the extraction of up to $s - 1$ lead summary sentences and an arbitrary number of supporting sentences.

Using training data, we obtained a maximum-likelihood estimate for each transition probability. This formed an estimate $\boldsymbol{M}$ for the transition matrix for our Markov chain, where element (i, j) of $\boldsymbol{M}$ was the estimated probability of transitioning from state i to state j.

Associated with each state i is an output function, $b_i(\mathbf{O}) = \Pr(\mathbf{O} \mid \text{state } i)$, where $\mathbf{O}$ is an observed feature vector. We made the simplifying assumption that the features are multivariate normal. To estimate each state's output function, we use the training data to compute the maximum-likelihood estimate of its mean and covariance matrix. We estimated $2s + 1$ means but assumed that all the output functions shared a common covariance matrix.

Thus, our model consists of three parts: p, the initial state distribution; $\boldsymbol{M}$, the Markov transition matrix; and B, the collection of multivariate normal distributions associated with each state.

Let $\alpha_t(i)$ be the probability that we have observed the sequence $\{\mathbf{O}_1, \mathbf{O}_2, \ldots, \mathbf{O}_t\}$ and are currently in state i $(1 \le i \le N)$ of our HMM. We can compute $\alpha_t(i)$ recursively as follows. Let $\alpha_1(i) = p(i)$ and compute

$$\alpha_t = D_{\mathbf{O}_t} \boldsymbol{M}^T \alpha_{t-1} \quad \text{for} \quad t = 2, \ldots, T,$$

where T is the number of sentences in the document and

$$D_{\mathbf{O}_t} = I - \operatorname{diag}\{b_1(o_1), b_2(o_2), \ldots, b_{2s+1}(o_{2s+1})\}.$$

I is the identity matrix and $b_j(o_i)$ is the cumulative density function for $o_i = (\mathbf{O}_t - \mu_i)^T \Sigma^{-1} (\mathbf{O}_t - \mu_i)$, where μ_i is the mean for the ith state and Σ is the covariance matrix. If $\mathbf{O}$ is multivariate normal, then the variable o has a χ^2 distribution, with the number of degrees of freedom equal to the number of components in $\mathbf{O}_i$.

The probability of the entire observation sequence given the model is given by

$$\omega \equiv \Pr(\mathbf{O}) = \sum_{i=1}^{2s+1} \alpha_T(i).$$

We define $\beta_t(i)$ to be the probability that we will observe the sequence $\{\mathbf{O}_{t+1}, \mathbf{O}_{t+2}, \ldots, \mathbf{O}_T\}$ given that we are at state i of our HMM. A backwards recursion lets us compute $\beta_t(i)$ by initializing β_T to all ones and then computing

$$\beta_t = \boldsymbol{M} D_{\mathbf{O}_{t+1}} \beta_{t+1} \quad \text{for} \quad t = T - 1, \ldots, 1.$$

We combine these two recursions to form $\gamma_t(i)$—the probability of being in state i for sentence t given the observation sequence $\{\mathbf{O}_1, \mathbf{O}_2, \ldots, \mathbf{O}_T\}$and the HMM. The formula is

$$\gamma_t(i) = (\alpha_t(i)\beta_t(i))/\omega.$$

These γs, computed using the "forward-backward" algorithm just mentioned, are used to determine the most likely states corresponding to the given observations. We compute the probability that a sentence is a summary sentence by summing $\gamma_t(i)$ over all even values of i. This posterior probability is used to select the most likely summary sentences. We denote this probability as

$$g_t = \sum_{i \text{ even}} \gamma_t(i).$$

The system chooses sentences by score, including the highest-scoring sentences until it meets the 100-word length or exceeds it by some constrained amount. The system then

NLP Features for Summarization

Researchers entered a variety of systems in the Document Understanding Conference (DUC 2002) competition. A noticeable difference from DUC 2001 was some systems' greater reliance on parsing and information extraction technology.

Our text summarization system retrieves relevant documents and produces indicative summaries from a huge, heterogeneous, and constantly changing set of documents. It does this retrieval and summarization in real time, and users typically want a summary containing four to five sentences. So, our environment shares many features with commercial environments and is somewhat different from that of the DUC experiments.

Applying the schema from *Advances in Automatic Text Summarization*,[1] we can characterize summarization approaches as surface-level, entity-level, and discourse-level. We somewhat adapt this schema for our discussion of NLP enhancements. Table A lists types of surface-level, entity-level, and discourse-level features that DUC 2002 systems exploited. The table includes a feature description and, wherever possible, at least one referent DUC 2002 system. We define status as

- Current—applied in our summarizer
- Explored—investigated yet abandoned
- Potential—feasible for our environment
- Promising—interesting but not feasible for our environment

An overview of this table reveals the constraints under which an operational system labors.

First, in sharp contrast to our approach, many DUC 2002 systems used entity-based features and discourse-level features, which were stages in modular natural-language-processing-enhanced systems. Many of these techniques lack portability and robustness.

Second, with few exceptions, efficient and effective syntactic parsers have not been inserted into real applications with volume and speed demands such as ours. Our system does incorporate named-entity recognition software, but templating techniques with manual tailoring for domain knowledge are only feasible in narrow domains and are not for high volumes of heterogeneous data.

Third, topic classification was prevalent in the DUC 2002 entries. Varying summarization strategies according to topic type was effective, but differentiating processing by document topic types would reduce the generality of our current approach and that of other commercial summarizers.

Table A. Surface-, entity- and discourse-level features. References denote example Document Understanding Conference II systems.

Feature	Description	Status
Surface level		
Thematic feature[2,3]	Statistically salient terms	Current
Location[4]	Sentence position in paragraph	Explored
Location	Sentence/paragraph position in text	Current
Background[5]	Headline	Potential
Background[6]	Sentence length	Current
Background	User query	Current
Cue words	Discourse markers	Current (heuristic)
Cue words[7,8]	Bonus terms, stigma terms	Promising
Entity level		
Similarity[9]	Vocabulary overlap	Potential
Proximity	Distance between text units	
Co-occurrence	Related words based on co-occurrence	Explored
Thesaural relationships[10]	Synonymy, hypernymy, part-of relations	Explored, potential
Coreference[11]	Referring expressions such as noun phrases	Promising
Syntactic relations[12,13]	Based on parse trees	Promising
Meaning representation[11,13]	Based on predicate-argument structure	Promising
Relations		
Discourse level		
Format of document[11–13]	Topic classification	Promising
Threads of topics[9,10,14]	Topic segmentation	Potential
Document discourse structure	Rhetorical structure	Explored, potential

reorders the selected sentences in their original document order to create the final summary.

Additionally, we applied two heuristics. The first heuristic identified and eliminated stereotypic, nonsummary sentences (boilerplate) before the HMM; the second removed discourse markers (And, But, and so on) occurring at the start of sentences that the HMM selected. Applying this second heuristic made our summaries *abstracts* rather than *extracts*.

Evaluation

Previous training of our text summarizer was based on using human-generated extract summaries. Because NIST provided no such data, we generated our own. Analysts mapped from the NIST-provided multidocument summaries to the information source sentences in each of 148 documents (half of the training data). We initially tried this using an automatic process—a cosine score—but abandoned it when we found that the automatically generated summaries had little content overlap with the NIST-generated summaries. We trained and tested the HMM using 119 of the tagged documents; we discarded the remaining 29 documents owing to problems with sentence boundaries. These training extracts were generally longer than the required 100-word count because the reference summaries often drew information from multiple sentences. We used the precision—as measured by the number of sentences in the extract that agreed with the human extract—as a simple score.

We performed a tenfold cross-validation using the 119 extracts. The precision for the 100-word single-document extracts was 0.55, an improvement over our precision of 0.52 for DUC 2001.[8] This gain was largely due to our use of cleaner training data. The boilerplate sentence recognition heuristic improved

Fourth, we have not applied thematic approaches that characterize the collection but have investigated the use of discourse structure. In DUC 2001, we explored using discourse parsing to generate more informative summaries[15] and learned that knowledge from manually built trees helps generate informative single-document summaries. However, implementation awaits further advances in both linguistic research in rhetorical structure theory and discourse-parsing technology.

Realistically, however, we can still strengthen our approach in some areas. Incorporating headlines, a surface-level background feature, would be useful. Examining similarity scores[5] or coreferencing with document named-entities[4] is feasible. Additionally, some combination of lexical chaining and topic segmentation seems realistic if we can address WordNet performance issues. Although lexical chaining[16] and topic segmentation[10] are actually used to construct the summaries, our goal would be to incorporate this knowledge as discourse-level features.

References

1. I. Mani and M. Maybury, eds., *Advances in Automatic Text Summarization*, MIT Press, Cambridge, Mass., 1999.
2. H. Zha and X. Ji, "Generating Extractive Summaries Using Mutual Reinforcement Principle and Sentence Clustering," *DUC 2002 Conf. Proc.*, Nat'l Inst. of Standards and Technology, Gaithersburg, Md., 2002; http://www-nlpir.nist.gov/projects/duc/pubs.html.
3. C.-Y. Lin and E. Hovy, "The Automatic Acquisition of Topic Signatures for Text Summarization," *Proc. 18th Int'l Conf. Computational Linguistics* (COLING 2000), 2000.
4. P. Lal and S. Ruger, "Extract-Based Summarization with Simplification," *DUC 2002 Conf. Proc.*, Nat'l Inst. of Standards and Technology, Gaithersburg, Md., 2002; http://www-nlpir.nist.gov/projects/duc/pubs.html.
5. T. Hirao et al., "Text Summarization System for DUC-2002," *DUC 2002 Conf. Proc.*, Nat'l Inst. of Standards and Technology, Gaithersburg, Md., 2002; http://www-nlpir.nist.gov/projects/duc/pubs.html.
6. H. van Halteren, "Writing Style Recognition and Sentence Extraction," *DUC 2002 Conf. Proc.*, Nat'l Inst. of Standards and Technology, Gaithersburg, Md., 2002; http://www-nlpir.nist.gov/projects/duc/pubs.html.
7. W. Kraaij, M. Spitters, and A. Hulth, "Headline Extraction Based on a Combination of Uni- and Multidocument Summarization Techniques," *DUC 2002 Conf. Proc.*, Nat'l Inst. of Standards and Technology, Gaithersburg, Md., 2002; http://www-nlpir.nist.gov/projects/duc/pubs.html.
8. C.-Y. Lin and E. Hovy, "NeATS in DUC 2002," *DUC 2002 Conf. Proc.*, Nat'l Inst. of Standards and Technology, Gaithersburg, Md., 2002; http://www-nlpir.nist.gov/projects/duc/pubs.html.
9. M. Karamuftuoglu, "An Approach to Summarization Based on Lexical Bonds," *DUC 2002 Conf. Proc.*, Nat'l Inst. of Standards and Technology, Gaithersburg, Md., 2002; http://www-nlpir.nist.gov/projects/duc/pubs.html.
10. R. Angheluta, R. DeBuser, and M.F. Moens, "The Use of Topic Segmentation for Automatic Summarization," *DUC 2002 Conf. Proc.*, Nat'l Inst. of Standards and Technology, Gaithersburg, Md., 2002; http://www-nlpir.nist.gov/projects/duc/pubs.html.
11. S. Harabigiu and F. Lacatusu, "Generating Single and Multi-document Summaries with GISTEXTER," *DUC 2002 Conf. Proc.*, Nat'l Inst. of Standards and Technology, Gaithersburg, Md., 2002; http://www-nlpir.nist.gov/projects/duc/pubs.html.
12. K. McKeown, R. Barzilay, and S. Blair-Goldensolhn, "The Columbia Multidocument Summarizer for DUC 2002," *DUC 2002 Conf. Proc.*, Nat'l Inst. of Standards and Technology, Gaithersburg, Md., 2002; http://www-nlpir.nist.gov/projects/duc/pubs.html.
13. D. Marcu, "Discourse Trees are Good Indicators of Importance in Text," *Advances in Text Summarization*, MIT Press, Cambridge, Mass., 1999, pp. 123–136.
14. M. Brunn, Y. Chali, and B. Dufour, "UofL Summarizer at DUC 2002," *DUC 2002 Conf. Proc.*, Nat'l Inst. of Standards and Technology, Gaithersburg, Md., 2002; http://www-nlpir.nist.gov/projects/duc/pubs.html.
15. L. Carlson et al., "An Empirical Study of the Relation between Abstracts, Extracts, and the Discourse Structure of Texts," *DUC 2001 Conf. Proc.*, Nat'l Inst. of Standards and Technology, Gaithersburg, Md., 2001; http://www-nlpir.nist.gov/projects/duc/pubs.html.
16. R. Barzilay and M. Elhadad, "Using Lexical Chains for Text Summarization," *Advances in Text Summarization*, MIT Press, Cambridge, Mass., 1999, pp. 111–121.

the precision further, increasing it to 0.57 for 100-word single-document summaries.

To verify that generating single-document summaries was useful for generating multidocument summaries, we explored the role that single-document summaries play when a human creates a multidocument summary. Using the SRA TagTool, an experienced analyst mapped the individual EDUs (elementary discourse units—for more information, see the "Further Reading" sidebar) in the (NIST-provided) 100-word multidocument reference summary to EDUs in the single-document reference summaries for five DUC 2001 training sets—d02, d16, d17, d25, and d35. We assumed that successfully mapping the multidocument reference summary's content to the single-document reference summaries in a set would suggest that our technique might be replicating an analytic strategy; that is, the writer would analyze what was important in each document, then synthesize the substantial content from the summaries to create the multidocument summary. Our annotation verified that the single-document summaries could have been used to generate many of the multidocument summaries and documented that the multidocument EDUs often referenced more than one single-document summary.

Cases exist, however, where the information in multidocument EDUs is not overtly referenced in the single-document summaries. In these, the summary author either relied on world knowledge, used inference, or used information in a document that did not appear in the single-document summaries. We discuss the implications of this in more detail later (see the "Shortcomings in Human Multidocument Summaries" section).

We also assessed our single-document summarization's coverage. Poor coverage

Table 1. *F*-scores for single-document summaries.

System	*F*-score[1]
sys27	0.475
sys19	0.469
sys28	0.441
sys15	0.435
sys31	0.433
sys29	0.432
sys21	0.430
sys23	0.408
sys18	0.368
sys25	0.368
sys16	0.363
sys31	0.148
sys17	0.138
base1[2]	0.466

1. Our *F*-score, *F*1, is *(2*precision*recall)/(precision + recall).*
2. Base1 was created by taking the first 100 white-space delimited, nontag tokens in the document.

here would mean poor multidocument coverage. The same analyst mapped the EDUs in the 100-word multidocument reference summary to EDUs in our machine-generated single-document summaries for the five DUC 2001 training sets previously mentioned. More than half of the multidocument human summary content was missing from our system-generated single-document summaries. Our system was not selecting the same information as the human.

Only one system (sys27) of the DUC 2002 participants significantly outscored the single-document summary baseline (base1), further confirming this finding (see Table 1). Our system (sys28) scored third but lower than the baseline. Yet we know from user studies that the summaries our system generates are far more useful to them than those the baseline method generates. We assume that this is also true for other systems.

Generating multidocument abstracts

We investigated two methods for multidocument summarization. Both used the HMM described in the single-document algorithm section to score each sentence in the document set using posterior probability. We then used the top-scoring sentences as candidates for the multidocument summary.

The algorithm

The HMM selects enough candidates to generate an extract of twice the maximum size requested, which for DUC would be 800 words. We use the candidate sentences to form a token-sentence matrix, ***A***, with each token (non-stop word) as a row in the matrix and each of the candidate sentences a column. (One could use a more linguistically complex construct in place of a token without changing our algorithm. Using a more complex structure is part of our ongoing research.) Using the simplistic definition that a token is important if it appears in a sentence, an element in ***A*** is 1 if the token appears in the sentence and 0 otherwise. We normalize ***A***'s columns so their 2-norm equals the posterior probability given by the HMM. We wish to choose columns (sentences) from ***A*** that cover the tokens well. We considered two approaches to solving this problem: pivoted QR factorization and a singular value decomposition method.[9] The latter method is considered more robust when ***A*** is ill conditioned, which might happen if the HMM selected several sentences that collectively had considerable overlap. However, we report only on the pivoted QR method.

Pivoted QR factorization attempts to select columns of ***A*** in the order of their importance in spanning the subspace spanned by all the columns. The algorithm initially rates sentences with a large norm as very important. At each stage, the algorithm selects as the next sentence the one with the largest norm. This choice is called *pivoting*.

Once a sentence is selected, the importance measures for the remaining sentences need updating because tokens shared with the selected sentence are no longer important to capture. The algorithm reduces the norm of each column by subtracting off the component of that column that lies in the direction of the column just added. This process of making the remaining matrix orthogonal to the previously chosen columns forms the *QR decomposition*.

The first *r* sentences that the pivoted QR method selects form the summary, with a choice of *r* that causes the summary length to be close to the target length. The standard implementation of the pivoted QR decomposition is a Gram-Schmidt process.[10]

The system outputs the sentences in their original document order. The document order is lexicographical, which results in a temporal order in a group of documents from the same source (for example, the *Wall Street Journal*) owing to the document file naming convention. We also applied the discourse and boilerplate heuristics before and after sentence selection, respectively.

Table 2. *F*-scores for all multidocument summaries.

System	10 words	50 words	100 words	200 words
sys19	0.827	0.489	0.475	0.451
sys26	0.664	0.548	0.418	0.423
sys28	—	0.439	0.423	0.384
sys24	—	0.318	0.387	0.393
sys20	0.489	0.314	0.369	0.388
sys29	0.376	0.380	0.320	0.322
sys25	0.445	0.325	0.315	0.326
sys16	0.454	0.278	0.297	0.265
base3[1]	—	0.421	0.434	0.413

1. Base3 comprises the first sentence in each consecutive document in the set, up to the summary. This was the higher scoring of the two available baselines.

Evaluation

Overall, DUC 2002 system performance on multidocument summarization was low. Only a single system (sys19) consistently beat the multidocument baseline (base3) in terms of agreement with a human-generated summary, with the second-best system (sys26) beating it in two out of three cases. Our own system (sys28) was third in the overall rankings, outperforming the baseline in one of the three cases. For multidocument summaries, each system's ranking in each of the 50-, 100-, and 200-word summaries (10-word summaries were omitted because not all systems participated) were added to determine the overall ranking. Table 2 shows multidocument summary results.

By examining the summaries together with the single-document summary input, we learned that our summaries lack coverage and topical focus. The system exploits only a few documents in the collection, routinely extracting two to three initial document sentences from only two to three of the documents.

For example, for DUC document set d25, our system-generated summary references source information from only four of the 12 documents in the collection. Our feature set leads to the selection of relatively long sentences. To increase our coverage, we need a technique for compressing selected sentences. Our summaries often inherit multiple summary focuses that characterize the individual documents' leads. We must also enhance our approach to capture and use the collection's thematic nature. We doubt that improving the individual summaries' thematic focus will lead to a thematic focus for the multidocument summary.

Figures 2 and 3 demonstrate the extremes in our algorithm's performance. Figure 2 shows the 200-word summary our system generated for DUC document set d100, which contained 15 articles about an unauthorized autobiography, a film, and some peripheral information about John Lennon (about 10 years after his death). The summary contains seven sentences, and each was selected from a different document. The *F*-score for this summary was 0—that is, no overlap existed for each of the reference (human-generated) summaries. Additionally, this summary was poorly organized, showed no continuity of ideas, and even generated the wrong conclusion about the author of the book referenced in the first sentence.

> A biography that portrays John Lennon as a drug-addled, anorexic bisexual who raged his way from Liverpool to New York City is "totally fiction," Yoko Ono said in a national radio broadcast. The image of a dour, shoeless English boy and his absent, carefree mother prompted Julia Baird and Geoffrey Giuliano to collaborate on a book. Thousands of John Lennon fans, some too young to remember the Beatles, swayed and sang along with songs including his anthem "Imagine" at a hometown tribute. Official newspapers, which once denounced the Beatles as harmful pests from the decadent West, joined a worldwide celebration and ran stories of tribute about the late John Lennon, who would have turned 50 on Tuesday. John Lennon's worldwide message of peace was delivered Tuesday as his song "Imagine" was played simultaneously for 1 billion people in 130 countries to celebrate what would have been his 50th birthday. The flaky avant-garde artist whose idea of music was to scream inside a plastic bag while John Lennon played guitar? Say what you want about Albert Goldman, the author of the new biography, "The Lives of John Lennon" (Morrow, $22.95), but you've got to hand it to him: This guy is one ambitious sleazemonger.

Figure 2. Our 200-word summary for document set d100—a bad summary.

> Security Council today swiftly condemned Iraq's invasion of Kuwait, demanding an unconditional withdrawal of Iraqi troops and calling for immediate negotiations between the countries. An emergency Arab League Council failed to publicly condemn Iraq's invasion of Kuwait on Thursday and adjourned its meeting for 24 hours as Arab leaders held urgent consultations on how to handle the crisis. THE INVASION: Iraq's troops, led by about 350 tanks, crossed the border at dawn Thursday, and seized the Kuwaiti palace and government buildings 40 miles away. Iraqi President Saddam Hussein seemed determined to solve his financial problems and fulfill territorial ambitions by dethroning the government of neighboring Kuwait. Saudi Arabia has been subdued in its reaction to Iraq's invasion of Kuwait, which places Iraqi forces near the oil-rich kingdom's border. UNITED NATIONS _ The United States urged other U.N. members at talks Sunday to impose broad economic and military sanctions on Iraq. Ruling party members said today that the government, fearing a U.S. air attack, was distributing automatic weapons to tens of thousands of supporters and preparing to evacuate the Iraqi capital. Security Council today overwhelmingly approved sweeping trade and military sanctions against Iraq, including a ban on oil purchases, to punish Baghdad for invading Kuwait.

Figure 3. Our 200-word summary for document set d110—a good summary.

Figure 3 shows the 200-word summary that we generated for DUC document set d110, which contained 15 articles about the Iraqi invasion of Kuwait. This summary contains eight sentences, each from a different document. In contrast to the other summary, this one has continuity and presents no misleading information. This summary's *F*-scores were 0.4286 and 0.5714, respectively, for the two reference summaries. Clearly, the type of document collection we are working with dramatically affects our summary quality.

Table 3. Elementary discourse unit distribution for the d25 data set.

Document	EDU-1	EDU-2	EDU-3	EDU-4	EDU-5	EDU-6	EDU-7	EDU-8	EDU-9	EDU-10	EDU-11	EDU-12
LA053089-0081	0	0	0	0	0	1	2	2	6	5	0	0
LA04i290-0125	0	0	0	0	0	0	0	0	0	0	0	0
LA092290-0175	0	0	0	0	0	0	0	0	0	0	1	0
LA111989-0125	1	1	2	2	3	0	0	0	2	2	1	5
LA112389-0104	0	0	0	0	1	0	0	0	2	2	0	0
LA113089-0118	0	0	0	0	2	0	0	0	0	0	0	0
LA121590-0056	0	0	0	0	0	1	1	0	3	2	0	0
SJMN91-06340029	0	0	0	0	0	0	0	0	0	0	1	1
WSJ900420-0022	0	0	0	0	0	0	0	0	0	0	0	2
WSJ911213-0029	0	0	0	0	0	0	0	0	0	0	2	0

Table 4. Elementary discourse units not subsumed per number of EDUs in the smaller abstract.

Document collection	50-word to 100-word	100-word to 200-word	200-word to 400-word
d02	1/4	1/8	4/22
d16	5/6	5/11	2/25
d17	0/0	0/0	2/27
d25	1/5	3/12	2/24
d35	0/0	0/0	0/0

Although we expect NLP enhancements to our text summarizer to improve performance at the single-document level, we believe that regularization of multidocument summary generation by summary authors will prove even more beneficial to the generation and evaluation of multidocument summaries.

Shortcomings in human multidocument summaries

We analyzed the information source in 23 human-referent summaries that the DUC training set provided. Using the SRA TagTool, an experienced analyst mapped each summary's individual EDUs to relevant sentences in the original documents. For each collection, we created a reference table that identifies the sources of summary information by gauging

- Popularity within a document—the frequency with which a summary EDU occurred in any given document
- Breadth across the collection—the number of documents within the collection in which the EDU source occurred

Table 3 shows EDU distribution for DUC set d25. Summary authors heavily utilized some documents and gave others little attention. Also, EDU sources can occur throughout the document body. Some EDUs are source-rich (EDU-10), and we can trace them to multiple documents. Others are source-poor (EDUs-1 and -2) and arise from a single sentence within the collection. Some EDUs (although none in this example) have no traceable text source in the document collection. This phenomenon was similar to our discovery that for some information in the single-document referent summaries, we could find no information source in the actual documents. These data and source referent charts should be useful for analyzing how the writers processed the collection.

We discovered considerable variation in coverage among different-length summaries that summarized the same collection. We analyzed the same five DUC 2001 training sets (d02, d16, d17, d25, and d35) and compared coverage for three summary pairings—50-word compared to 100-word, 100-word compared to 200-word, and 200-word compared to 400-word—for the five sets. Specifically, for each pairing, an experienced analyst judged whether the EDUs in the shorter summary occurred in the longer summary. We assumed that a 50-word summary would represent the most essential information in the collection and that in lengthening the summary the writer would only add information. However, the summary writers frequently "lost" data that we assumed would be subsumed into the longer summary. Table 4 shows, for example, that for data set d16, five of six EDUs in the 50-word summary are not subsumed in the 100-word summary.

These two studies on information sources and lengthening in multidocument summaries showed that considerable latitude existed in summary creation. Each document set had three human summaries, so we compared them to confirm what appeared to be an extreme variation in human performance.

For each of the 15 DUC test sets, we compared the three summaries. For each summary we divided the number of EDUs in the summary that were unique (that is, not contained in either of the other two summaries) by the total number of EDUs in the summary. This gave us 45 scores. Their median value was 0.60—for a typical summary, 60 percent of the EDUs were not found in either of the other two summaries for the same document set. Therefore, assuming the EDUs indicate content, the majority of content in one summary is distinct from the content in the other two summaries. This suggests that the summary author greatly influences content and that any one multidocument summary might not provide a representative description of the original document set.

We can visualize further evidence of this author variation by comparing the discourse trees for the summaries of three authors summarizing the same collection. An original annotator of the RST-Corpus (see the "Further Reading" sidebar) applied the Rhetorical Structure Theory framework and created discourse trees for both 50-word and 100-word summaries for five DUC test sets (30 total). These discourse trees capture the multidocument summary's rhetorical structure for each author. They depict the relations between text spans and the centrality of information with the concepts of nuclearity and satellite. They

Further Reading

Hidden Markov models

L.E. Baum et al., "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains," *Annals of Mathematical Statistics*, vol. 41, no. 1, 1970, pp. 164–171.

L.R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," *Proc. IEEE*, vol. 77, no. 2, 1989, pp. 257–285.

Query-driven summaries

J.D. Schlesinger, D.J. Baker, and R.L. Donaway, "Using Document Features and Statistical Modeling to Improve Query-Based Summarization," *DUC 2001 Conf. Proc.*, Nat'l Inst. of Standards and Technology, Gaithersburg, Md., 2001; http:// www-nlpir.nist.gov/projects/duc/pubs.html.

Elementary discourse units

L. Carlson, D. Marcu, and M.E. Okurowski, "Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory," to be published in *Discourse and Dialogues*, Kluwer Academic Press, Dordrecht, Netherlands, 2003.

let us visualize each author's individual composition strategy and help us understand their language processing strategies.

We compared the three discourse trees for each of the five sets and observed that the trees for two collections had richer discourse structures characterized by multiple types of different discourse relations while trees in the other three sets were simple list structures. We contend that the content of some document collections leads the summary author to adopt a particular rhetorical structure. We posit that a collection, like a document, might have a rhetorical structure—a kind of "superstructure." The content of the original documents in the collection, as a whole, has a coherent structure. The d39 (Chunnel construction) and d37 (worldwide assassination) collections are good examples of this. The three discourse trees for the 50-word summaries for d39, which Figure 4 shows, demonstrate how this superstructure affects which information is selected for the multidocument summary. The three authors elaborate on the Chunnel construction with a contrastive relation that compares French and British receptivity. Figures 4a and 4b both contain satellite consequence relationships on the positive impact on travel time. All three trees depict the saliency of the difference between how France and Britain view the Chunnel. It appears that when a collection of documents has a rhetorical superstructure, the content of the multidocument summaries written by different authors will be more likely to overlap.

However, given this rhetorical structure, the individual authors still apply their own standards of saliency. Table 5 shows the actual content of the contrastive relationship. We can see that the uniformity the superstructure imposes has limits.

Moreover, when collections as a whole lack any rhetorical superstructure, individual authors have no framework upon which to build their summaries. The collection and the authors' corresponding discourse trees lack rhetorical specificity. This leads to discourse structures that have list structures composed of document facts. Discourse trees for document sets d24 (Elizabeth Taylor), d28 (marathons), and d11 (tornadoes) for the 50-word summary are characterized by this phenomenon and have low content overlap. In the absence of rhetorical guidance, the authors' individual standards of saliency determine what document facts are compiled in the lists.

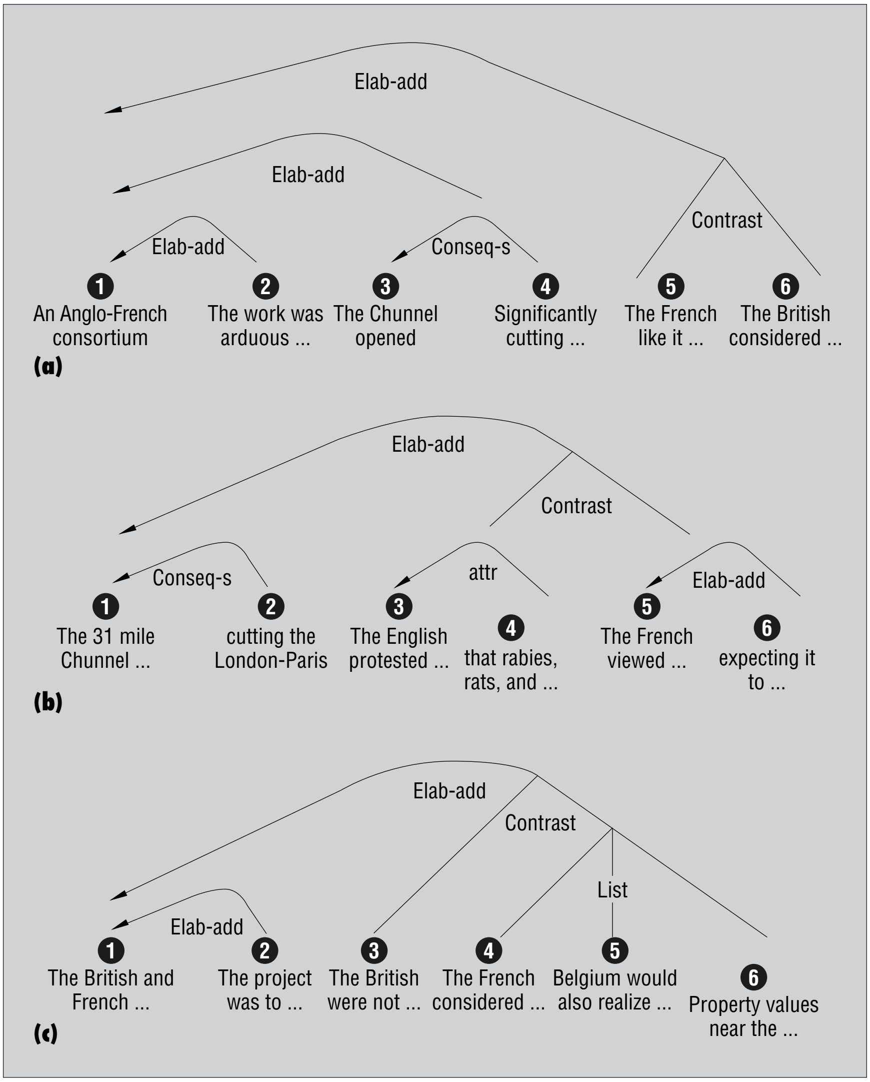
Elab-add
Elab-add
Contrast
Elab-add
Conseq-s
1 An Anglo-French consortium
2 The work was arduous ...
3 The Chunnel opened
4 Significantly cutting ...
5 The French like it ...
6 The British considered ...
(a)
Elab-add
Contrast
Conseq-s
attr
Elab-add
1 The 31 mile Chunnel ...
2 cutting the London-Paris
3 The English protested ...
4 that rabies, rats, and ...
5 The French viewed ...
6 expecting it to ...
(b)
Elab-add
Contrast
List
Elab-add
1 The British and French ...
2 The project was to ...
3 The British were not ...
4 The French considered ...
5 Belgium would also realize ...
6 Property values near the ...
(c)

Figure 4. Discourse trees for (a) author 1, (b) author 2, and (c) author 3.

In DUC 2002, Kathleen McKeown and her colleagues[11] argued that the selection strategy for the DUC collections affected the redundancy typically characteristic of data pulls. They called for more effective categorization of the data into clusters. We acknowledge the role that redundancy plays, but believe that in addition to redundancy, the presence or absence of a collection superstructure can affect content saliency. We

Table 5. Comparison for contrastive relationship.

	Contrast	
Summary	**British**	**French**
Figure 4a	The British considered it a threat to their island identity.	The French liked it.
Figure 4b	The English initially protested that rabies, rats, and terrorists would come through.	The French viewed the Chunnel positively, expecting it to revitalize their depressed northern region.
Figure 4c	The British were not eager partners.	The French, however, considered it an economic boost.

would support a more task-oriented evaluation, but we also strongly advocate for summarization guidelines and procedures. We surmise that the authors adopt their own idiosyncratic methodologies to create a rhetorical superstructure for collections that lack "rhetorical support." Additionally, even when rhetorical superstructure is available, the lack of summarization procedures for identifying and gauging what is salient contributes to "overriding" the existing superstructure. In some sense, it would be analogous to each annotator of a named-entity corpus defining his or her own concept of what a name or organization is. Another analogy is that of an annotator of a syntactic treebank corpus parsing sentences with his or her own personally defined formalisms. Because summarization is an even more poorly understood complex cognitive task, we recommend creating a methodology with guidelines based on sound research into the summary authors' language processing.

Current systems for multidocument summarization produce summaries that differ greatly from each other and from human referent summaries. Unfortunately, human summaries show a similar degree of variability. There are many potential new research directions that the community can take for multidocument summarization. To assess the value of these directions, though, we need access to high-quality annotated data. Therefore, one of the most beneficial directions for the community would be developing guidelines for consistent generation of training and test corpora of multidocument summaries.

References

1. I. Mani and M. Maybury, eds., *Advances in Automatic Text Summarization*, MIT Press, Cambridge, Mass., 1999.
2. H. Borko and S. Chatman, "Criteria for Acceptable Abstracts: A Survey of Abstractors' Instructions," *Am. Documentation*, vol. 14, no. 2, 1970, pp. 149–160.
3. R. Cremmins, *The Art of Abstracting*, Information Resources Press, Arlington, Va., 1996.
4. P. Over, "Overview of DUC 2002," *DUC 2002 Conf. Proc.*, Nat'l Inst. of Standards and Technology, Gaithersburg, Md., 2002; http://www-nlpir.nist.gov/projects/duc/pubs.html.
5. J. Kupiec, J. Pedersen, and F. Chen. "A Trainable Document Summarizer," *Proc. 18th Ann. Int'l SIGIR Conf. Research and Development in Information Retrieval*, ACM Press, New York, 1995, pp. 68–73.
6. C. Aone et al., "A Scalable Summarization System Using Robust NLP," *Proc. ACL'97/EACL'97 Workshop Intelligent Scalable Text Summarization*, Assoc. Computational Linguistics, East Stroudsburg, Pa., 1997, pp. 66–73.
7. T. Dunning, "Accurate Methods for Statistics of Surprise and Coincidence," *Computational Linguistics*, vol. 19, no. 1, 1993, pp. 61–74.
8. J.M. Conroy et al., "Using HMM and Logistic Regression to Generate Extract Summaries for DUC," *DUC 2001 Conf. Proc.*, Nat'l Inst. of Standards and Technology, Gaithersburg, Md., 2001; http://www-nlpir.nist.gov/projects/duc/pubs.html.
9. G.H. Golub, V. Klema, and G.W. Stewart, *Rank Degeneracy and Least Squares Problems*, tech. report TR-456, Dept. Computer Science, Univ. of Maryland, College Park, Md., 1976.
10. J.M. Conroy and D.P. O'Leary, *Text Summarization via Hidden Markov Models and Pivoted QR Matrix Decomposition*, tech. report TR-4221, Dept. Computer Science, Univ. of Maryland, College Park, Md., 2001.
11. K. McKeown, R. Barzilay, and S. Blair-Goldensolhn, "The Columbia Multidocument Summarizer for DUC 2002," *DUC 2002 Conf. Proc.*, Nat'l Inst. of Standards and Technology, Gaithersburg, Md., 2002; http://www-nlpir.nist.gov/projects/duc/pubs.html.

The Authors

Judith D. Schlesinger is a research staff member at the IDA Center for Computing Sciences in Bowie, Md. Her research interests span programming and natural languages, and she has published papers on intelligent tutoring systems, parallel programming, and automatic summarization. She received her BS in math from Brooklyn College, CUNY/SUNY, her MS in computer and information science from Ohio State University, and her PhD in computer science from Johns Hopkins University. She is a member of the ACM, the IEEE Computer Society, and the Association for Computational Linguistics. Contact her at the Center for Computing Sciences, 17100 Science Dr., Bowie, MD 20715-4300; judith@super.org.

John M. Conroy is a research staff member for the IDA Center for Computing Sciences in Bowie, Md. His research interest is applications of numerical linear algebra. He received his PhD from the Applied Mathematics Program at the University of Maryland. He is a member of the Society for Industrial and Applied Mathematics, the IEEE, the IEEE Computer Society, and the Association for Computational Linguistics. Contact him at the Center for Computing Sciences, 17100 Science Dr., Bowie, MD 20715-4300; conroy@super.org.

Mary Ellen Okurowski is a senior researcher at the Department of Defense. She has been active in the human-language-technology arena for more than 15 years. She received MAs in Chinese language and literature and in linguistics from the University of Kansas and her PhD from Georgetown University. She is a member of the Association for Computational Linguistics. Contact her at meokuro@nsa.gov.

Dianne P. O'Leary is a professor in the Computer Science Department and Institute for Advanced Computer Studies at the University of Maryland. She received her BS in mathematics from Purdue University and her PhD in computer science from Stanford. She has authored over 60 journal articles on computational linear algebra and optimization, algorithms for high-performance computers, numerical solution of ill-posed problems, and scientific computing. She is a member of the ACM, SIAM, and the Association for Women in Mathematics. Contact her at the Computer Science Dept., Univ. of Maryland, College Park, MD 20742; oleary@cs.umd.edu; www.cs.umd.edu/users/oleary.