Using Leading Text for News Summaries: Evaluation Results and Implications for Commercial Summarization Applications

Mark Wasson
LEXIS-NEXIS, a Division of Reed Elsevier plc
9443 Springboro Pike
Miamisburg, Ohio 45342 USA
mark.wasson@lexis-nexis.com

Abstract

Leading text extracts created to support some online Boolean retrieval goals are evaluated for their acceptability as news document summaries. Results are presented and discussed from the perspective of commercial summarization technology needs.

1 Introduction

The Searchable LEAD system creates a Boolean query aid that helps some online customers limit their queries to the leading text of news documents. Customers who limit their Boolean queries to leading text usually see better precision and an increased emphasis on documents with major references to their topics in their retrieval results.

A research team investigating a sentence extraction approach to news summarization modified Searchable LEAD to create leading text extracts to use in a comparison between approaches. Leading text extracts had a much higher rate of acceptability as summaries than the team expected. Because that test was limited to 250 documents, we were not certain how well leading text would rate as summaries on a larger scale, such as across the NEXIS® news database. We also could not make any conclusive statements about where leading text extracts routinely fail as summaries for news documents.

This paper presents the results of an investigation into how Searchable LEAD-based leading text extracts rate as summaries and where those extracts fail. The results support the use of leading text as general purpose summaries for news documents.

2 Searchable LEAD Overview

Searchable LEAD was originally implemented to provide LEXIS-NEXIS customers with the means to limit Boolean queries to key parts of news documents. It is based on the premise that major entities and topics of news stories are usually introduced in the leading portion of news documents. Searchable LEAD targets the subset of news information customers who want to retrieve documents that contain major references to their targeted topics but not documents that only mention those topics in passing. These customers generally can expect higher precision and lower recall when they restrict their Boolean queries to the headline and leading text than if they were to apply their queries to the full text.

Documents in our news database have several text fields including HEADLINE and BODY fields. Searchable LEAD software identifies the leading portion of the BODY field and labels it the "LEAD" field. The amount of the BODY field that is included in the LEAD is based on document length. Minimum thresholds for the number of words, sentences and paragraphs to include in LEADs increase as document length increases.

In an examination of more than 9,000 news documents from more than 250 publications, we found that short documents usually begin with good topic-summarizing leading sentences, what we call the *logical lead*. Longer documents, however, more often begin with anecdotal information before presenting the logical lead. LEAD fields must be longer for these documents in order to include the logical lead in most instances. Using a fixed amount of leading text regardless of document length would have resulted in LEADs that include

too much text beyond the logical lead for shorter documents, and LEADs that miss the logical lead entirely for longer documents.

Customers can limit part or all of a Boolean query to the LEAD, as the following query shows:

LEAD(CLINTON) AND BUDGET

This query will retrieve only those documents that contain "Clinton" in the LEAD and "budget" anywhere in the document. Customers who use LEAD routinely combine it with the HEADLINE field.

We tested 20 queries on a database that contains 20 million documents from more than 10,000 English language news publications. Each query was applied to the HEADLINE and BODY fields (abbreviated here as HBODY) and to the HEADLINE and LEAD fields (HLEAD). Queries were limited by date in order to reduce the magnitude of the evaluation task. In order to obtain a more complete picture of recall, other queries were used to identify relevant documents that the tested queries missed. The results in Table 1 show that limiting Boolean queries to leading text can help Searchable LEAD's targeted customers.

Document Fields	Average Precision	Average Recall	Average of F-measures
HLEAD	.472	.600	.432
HBODY	.208	.793	.288

Table 1. Impact of LEAD Restrictions on Boolean Retrieval Quality (20-query test)

Searchable LEAD document processing software consists of a 500-statement PL1 program and a 23-rule sentence and paragraph boundary recognition grammar, and operates in a mainframe MVS environment. Searchable LEAD processes over 500,000 characters (90 news documents) per CPU second.

3 Related Work

There is a growing body of research into approaches for generating text summaries, including approaches based on sentence extraction (Kupiec et al., 1995), text generation from templates (McKeown and Radev, 1995) and machine-assisted abstraction (Tsou et al., 1992). Brandow et al. (1995) reported on a sentence extraction approach called the Automatic News Extraction System, or ANES. ANES combined statistical corpus analysis, signature word selection and sentence weighting to select sentences for inclusion in summaries. By varying the number of sentences selected, ANES-generated extracts could meet targeted summary lengths.

ANES was evaluated using a corpus of 250 documents from newswire, magazine and newspaper publications. ANES was used to generate three summaries for each document, targeting summary lengths of 60, 150 and 250 words. For a baseline comparison, a modified version of the Searchable LEAD software was used to create three fixed length leading text summaries for each document, also targeting lengths of 60, 150 and 250 words.

News analysts read each document and its corresponding summaries, and rated the summaries on their acceptability. Table 2 shows the results for each approach. Overall, 74% of the ANES summaries were judged to be acceptable. Unexpectedly, the acceptability rate for leading text summaries was significantly higher. Overall, 92% of the leading text summaries were judged to be acceptable.

Summary Length	ANES Acceptable	Leading Text Acceptable
60 words	68%	87%
150 words	76%	93%
250 words	78%	96%
Overall	74%	92%

Table 2. Acceptability Rates Comparison between ANES and Leading Text

The results for both approaches showed a promising start towards the goal of creating summaries for news documents. However, those results also raised questions about leading text. We wanted to better understand the value of leading texts as general purpose news document summaries.

4 Methodology

Our investigation had two goals: to verify on a larger scale the results that Brandow et al. (1995) suggested for leading text, and to determine whether there are easily definable indicators of where leading text extracts fare poorly as general purpose news document summaries.

We used the Searchable LEAD definition of leading text as our summaries. The LEAD fields vary in length based on overall document length, which we believe helps them capture the logical lead. Also, LEAD fields already existed in our news documents in support of another application, Boolean retrieval. We did not modify Searchable LEAD software or any LEAD fields for this investigation.

The test corpus consisted of 2,727 documents from more than 100 English language news publications. Documents were retrieved from our news database using several queries. Some queries were biased towards longer documents or to sources that provide transcripts. We believed that LEADs for such documents would pose more problems than would LEADs for typical news stories, based on past informal observations of LEAD fields. Because of the query bias, the test corpus does not represent our news database. For example, only 5.5% of the documents in the test corpus were less than 120 words long, whereas 18% of the documents in our news database are that short. Newspapers provide almost 60% of the documents in our news database but only a third of the test corpus documents.

In order to investigate where LEADs might fail as summaries, we assigned attributes to each document that allowed us to examine various subsets of the test corpus. Attributes included the following:

- BODY field and LEAD field word counts
- Source type (newspaper, wire service, newsletter, magazine, transcript service)
- Subject matter (biographical, financial, legal, legal news, other news, reviews, scientific)
- Document type (general news, which includes standard news articles, graphics, editorials, LEAD=BODY, letters/Q&A columns, and music and book reviews; lists; newsbriefs; and television program transcripts)
- United States or non-United States source

News analysts read each document and rated its corresponding LEAD field on its acceptability as a general purpose summary for that document. They rated the LEADs as either acceptable or unacceptable. Ratings were linked to document attributes in an evaluation file that contained one record for each document. This file was analyzed to obtain descriptive information about the test corpus and to compare attributes and ratings.

5 Results

Overall, 82.3% of LEADs were rated acceptable as summaries. However, because of differences between test corpus content and the content of our news database, this acceptability rate is not an overall indicator for our news database.

Document type was the most distinguishing attribute for identifying potential problem LEADs. For the general news document type, 94.1% of LEADs were rated acceptable as summaries. Acceptability rates were much lower for lists, newsbriefs and transcripts, as Table 3 shows.

Document Type	Number of Documents	Acceptability Rate
General News	1,951	94.1%
Lists	86	12.8%
Newsbriefs	191	24.6%
Transcripts	499	70.3%

Table 3. Acceptability Rates for Document Types

The 94.1% acceptability rate for general news documents is not appreciably different from the 92% average that Brandow et al. (1995) reported.

The results for lists and newsbriefs were not surprising. Such documents seldom have logical leads. Lists primarily consist of several like items, such as products and their prices, or companies and corresponding stock quotes. In rare instances, the BODY of a list type document includes a brief description of the contents of the list that Searchable LEAD can capture. In most cases, however, there is nothing meaningful for any technology to extract.

Newsbrief documents usually consist of several often unrelated stories combined into one document. In some newsbrief documents, however, there is an introduction that Searchable LEAD can exploit. This was especially true for newsbrief documents from wires (67.4% acceptability on 46 documents), but rarely true for either magazines (13.8% accept-

ability on 109 documents) or newspapers (3.1% acceptability on 32 documents).

LEADs for transcript type documents fared somewhat better, with source being a factor for these also. LEADs for transcripts from transcript sources were less likely to be rated acceptable (67.8% acceptability on 435 documents) than those from wires (90.0% acceptability on 40 documents) or newsletters (83.3% acceptability on 24 documents).

Among general news documents, only LEADs for the review sub-type had a low acceptability rate, as Table 4 shows.

General News Sub-types	Number of Documents	Acceptability Rate
News Articles	1,806	95.5%
Reviews	58	48.3%
All Others	87	95.4%

Table 4. Acceptability Rates for General News Document Sub-types

The distribution of list, newsbrief and transcript type documents was often the cause of other apparent problem-indicating attributes. For example, the overall acceptability rate for LEADs for United States sources was 80.1% on 2,141 documents, whereas the overall acceptability rate for non-United States sources was 90.4% on 586 documents. When list, newsbrief and transcript documents were removed, the acceptability rate for United States sources was 94.5% on 1,391 documents, and the acceptability rate for non-United States sources was 93.0% on 560 documents.

When examining other general news document attributes, we found that only LEADs for magazines had a somewhat lower acceptability rate (Table 5).

Source Type	Number of Documents	Acceptability Rate
Magazines	470	88.5%
Newsletters	217	98.2%
Newspapers	880	94.2%
Transcripts	7	100.0%
Wires	377	98.1%

Table 5. Acceptability Rates for General News by Source Type

The review sub-type was a factor here. Many of those were from magazines. Excluding those, the acceptability rate for magazine LEADs climbed to 92.5%, still lower than for any other source.

Document length was a factor for LEAD acceptability for the entire test corpus, but list, newsbrief and transcript type documents are typically longer than general news documents. Document length was less of a factor when looking only at LEADs for general news documents (Table 6).

BODY Length	Number of Documents	Acceptability Rate
0-119 words	151	97.4%
120-299 words	168	98.2%
300-599 words	312	95.8%
600-1199 words	548	94.9%
1200+ words	772	91.2%

Table 6. Acceptability Rates for General News by Document Length

The length of the LEAD itself was not tied to acceptability for either the entire test corpus or the general news document subset.

6 Discussion

The results of this investigation show that leading text can provide acceptable summaries for most general news documents. These results are consistent with Brandow et al. (1995). However, we also found that leading text is much less likely to provide acceptable summaries for news documents with certain structures, including list, newsbrief and transcript documents. We identified review type documents as a problem area, but these represent a small fraction of news data. More noteworthy was the observation that almost one of every eight LEADs for general news documents from magazines were rated unacceptable as summaries.

Kupiec et al. (1995) compared their trainable document summarizer results and similar amounts of leading text to manually constructed keys. The sentences that their summarizer extracted overlapped 42% of the sentences in the keys, compared to 24% for leading text. Both percentages are much lower than what Brandow et al. (1995) reported, but differences between the evaluation approaches used

is the probable reason. Kupiec et al. (1995) noted that there may be more than one good summary for a given document, something that a key approach to evaluation does not capture. Brandow et al. (1995) found this to be the case for some documents where all ANES-generated and leading text summaries were rated as acceptable. Some differences in results may also be attributed to the test data used. Kupiec et al. (1995) used scientific and technical documents rather than general news.

Leading text extracts such as the LEAD field are appealing for commercial use as summaries for a number of reasons. For general news documents, they are usually acceptable as summaries. They are easy and inexpensive to create. Leading text extracts also have two less obvious advantages over other approaches. First, legal restrictions often prevent us from manipulating copyrighted material. Leading text extracts often preserve the existing copyright. Second, when leading text fails as a summary, customers can see why. Customer understanding of how a data feature is created is often key to customer acceptance of that feature.

There are, however, a number of reasons why we need to consider alternatives to leading text. First, not all documents have a logical lead that can be exploited. In this investigation, we found that to be the case for most list and newsbrief documents and for many transcripts. Beyond news data, this holds for case law documents, many types of financial documents, and others.

Second, a static summary such as one based on leading text represents a "one size fits all" approach to summarization. Readers bring their own interests to documents. A dynamic summary generator, perhaps using readers' queries to guide it, can help readers focus on those parts of a document that are most relevant to them.

Third, a hybrid approach to summary generation may improve acceptability for news documents. Lin & Hovy (1997) describe methods for identifying the likely locations of topic-bearing sentences. Comparing the content of leading text extracts to predictions of topic-bearing sentences may help us predict where leading text fails as a summary so that we can direct more sophisticated approaches to those documents.

The commercial use of leading text summaries such as Searchable LEAD by no means suggests that news summarization is a solved problem. There are a number of data types where leading text has diminished or no value as a summary. Where it does succeed, an approach like Searchable LEAD may serve as a starting point for improved leading text summaries or as a benchmark for comparing alternatives that are not restricted by the limits inherent in leading text approaches to summarization.

Acknowledgments

I would like to thank colleagues Christi Wilson and David Schmeer for reading, rating and assigning attributes to the 2,727 test corpus documents and LEADs. I would also like to thank Breck Baldwin, Afsar Parhizgar, David Schmeer and Paul Zhang for their comments and suggestions.

References

Brandow, R., Mitze, K. and Rau, L. (1995) *Automatic Condensation of Electronic Publications by Sentence Selection*. Information Processing & Management, 31/5, pp. 675-685.

Kupiec, J., Pedersen, J., and Chen, F. (1995) *A Trainable Document Summarizer*. Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 68-73.

Lin, C.-Y., and Hovy, E. (1997) *Identifying Topics by Position*. Proceedings of the Fifth Conference on Applied Natural Language Processing, pp. 283-290.

McKeown, K., and Radev, D. (1995). *Generating Summaries of Multiple News Articles*. Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 74-82.

Tsou, B., Ho, H.-C., Lai, T., Lun, C., and Lin, H.-L. (1992) *A Knowledge-based Machine-aided System for Chinese Text Abstraction*. COLING-92 Proceedings, pp. 1039-1042.