Lexical Cohesion, Discourse Segmentation and Document Summarization

Branimir K. Boguraev and Mary S. Neff
IBM T.J. Watson Research Center, P.O. Box 704, Yorktown Heights, NY 10598, USA
bkb@watson.ibm.com, neff@watson.ibm.com

Abstract

Summaries automatically derived by sentence extraction are known to exhibit some coherence degradation, readability deterioration, and topical under-representation. We propose a strategy for improving upon these problems, aiming to generate more cohesive summaries by analyzing the lexical cohesion factors in the source document texts. As an initial experiment, we have looked at one particular factor, lexical repetition, which is instrumental to the topical make-up of a text. We have developed a framework for integrating a lexical repetition-based model of discourse segmentation capable of detecting shifts in topic, with a linguistically-aware summarizer which utilizes notions of salience and dynamically-adjustable size of the resulting summaries. We show that even by utilizing lexical repetition alone, summaries are of comparable, and under certain conditions better, quality than those delivered by a state-of-the-art sentence-based summarizer. This is encouraging for a broad platform of research which seeks to position a framework for the recognition and use of a number of cohesive devices in text as instrumental in the development of a wide range of content characterisation and document management tasks.

1 Introduction

As critical technology for managing information dissemination and use, document summarization has become the subject of active research (Mani and Maybury, 1999). However, with no coherent theory of summarization, and with no rigorous computational model of the summarization process, it is very hard to focus on specific aspects of it with the goal of improving the overall performance of a summarization system. Without knowing what factors contribute, and how, to the quality of a summary, the question of targeting just the right factors becomes one of propitious intuition, rather than informed judgement.

Still, a variety of strategies have been proposed to alleviate some of the problems associated with the summarization-by-sentence-extraction model. This paper elaborates a particular intuition which seems to have remained largely unexplored till now: the notion is to exploit some of the factors contributing to the coherence of the source text by carrying their effects over to the text's summary.

Summaries generated as a concatenation of sentences extracted (by some method) from the original document are, typically, sub-optimal. Of particular relevance to this research are the related problems of *coherence degradation*, *readability deterioration*, and *topical under-representation* (Boguraev et al., 1998). In essence, the deletion of arbitrary amount of source material between two sentences which end up adjacent in the summary has the potential of losing essential information. Examples like 'dangling' anaphors, whose antecedents have been lost, have been cited often enough; simple strategies like including the immediately preceding sentence in the summary have some effect. Still, these are simple strategies, prone to misfiring; moreover, other effects like the reversal of a core premise in an argument, or the introduction, and subsequent elaboration, of a new topic, are not easily handled by similar heuristics.

In an earlier work, we have argued that one way of addressing problems arising from sentence deletion would be to eschew the notion that sentences are suitable units for representing salient document highlights. Instead, we proposed that phrasal units of certain type, presented to retain strong ties with their context in the source, offer a characterization (as an abstraction) of the document content, which is

more sensitive to topical shifts than a sentence-based summary (Boguraev, Bellamy, and Kennedy, 1999). While such an approach clearly focuses on the topical under-representation problem, it deliberately compromises on readability[1], and thus produces document abstractions which are less than coherent. As we argue, there are situations where this may be acceptable.

In our current research, however, we start from the position that the underlying summarization technology is that of sentence-based, rather than phrase-based, extraction. Addressing the problems listed above now becomes a matter of somehow introducing a mechanism for independently assessing, and using, the *degree of cohesion* between individual sentences in the source document. Having some notion of how these sentences map onto the underlying themes (major topics) in the document becomes equally important.

1.1 Lexical cohesion

Authors use of a large number of rhetorical devices to make documents coherent. Our intuition is that by analysing such devices—or at the very least by being sensitive to their manifestation and interplay in a text—we can bring a moderately refined degree of discourse awareness into the summarization process. In the absence of full (or even partial) text understanding, we attempt to leverage a formalized notion of text cohesion, in particular *lexical cohesion*.

A number of linguistic studies focus on the operation and interaction of various *cohesive ties* (Halliday and Hasan, 1976), thus accounting for certain properties of the overall organization of a text discourse. More recently, Winter (1979), Phillips (1985), Phillips (1989) have looked at the devices that enforce lexical relationships and connect a discourse fragment with other discourse fragments.

Several key points emerge from this research. Cohesion can be best explained by focusing on how repetition is manifested, in numerous ways, across pairs of sentences. Repetition itself carries informational value, to the extent that it provides a reference point for interpreting what has changed (and thus, what is at the focus of attention of the discourse). This clearly goes beyond the simple notion that discourse fragments with shared content will also share vocabulary. As Phillips (1985) points out, the lexical inventory of a text is tightly organized in terms of collocation. It is this particular property that offers a handle on the overall organization of text, in general, and on the identification of *topic introduction* and *topic closure*, in particular.

At the same time, it turns out that just by being sensitive to patterns of repetition it is possible to distil *lexical chains* (Morris and Hirst, 1991), which are themselves representative of topic indicators. The interplay of repetition, lexical cohesion, and topicality thus facilitates a multi-perspective approach to determining relatedness among sentences (or other discourse fragments).

A variety of linguistic devices act as vehicles for repetition: at the level of interaction between words and phrases in the text, these include *lexical repetition*, *textual substitution* and the use of a range of *lexical relations*, *co-reference* and *ellipsis*, *paraphrasing*, *conjunction*, and so forth. A uniform approach to exploiting cohesive ties for improving coherence of summaries would aim to derive chains of sentences which focus on aspects of discourse entities or events; this would, in itself, require fairly detailed level of linguistic analysis of phenomena like synonymy, co-reference, ellipsis, and so forth (Hoey, 1991). While challenging, this is certainly possible. Fine-grained methods for computational analysis of cohesive ties in a text have been developed: see, for instance, (Fellbaum, 1999; Kennedy and Boguraev, 1996; Keller, 1994). In this paper, however, we focus on the extent to which lexical repetition alone can be used to determine the degree of cohesion between pairs of sentences.

[1]Lists of phrases, no matter how salient, are far from coherent prose; this is why the issue of appropriate presentation metaphors for such topical phrases is an important one, and is discussed at length in (Boguraev, Bellamy, and Kennedy, 1999).

Uniquely, this kind of intuition is explicitly addressed in recent work by Barzilay and Elhadad (1999). They introduce the notion of "lexical chains", derived by grouping together items related by repetition and certain lexical relations calculated by reference to the WORDNET lexical database (Fellbaum, 1999). A sequence of items in a chain highlights a discussion focused on a topic related to item(s) in the chain; using a metric for scoring chains, the more topical ones are selected; these are then taken as the basis of sentence extraction heuristics.

Even if Barzilay and Elhadad are frustrated by the high degree of polysemy of the words in WORDNET (not to mention the problems of coverage that WORDNET does, or rather does not, offer for more specialized domains), they report that in an intrinsic evaluation against human-constructed summaries, the system outperformed at least one commercial summarizer. This suggests that a purely lexical chains-based approach has potential worth investigating.

1.2 Discourse segmentation and sentence-based summarization

The starting point for us is a sentence-based summarizer which already has high scores in a discipline-wide evaluation initiative (Hand and Sundheim, 1998; Mani, Firmin, and Sundheim, 1999). We are thus seeking to improve upon what is already (by some measure; see 4.1 below) a good performance. Our strategy is to exploit lexical chaining specifically for the purposes of ensuring that the topics highlighted by any given chain are represented in the final summary.

We develop an operational definition of linear[2] discourse segmentation, where segments in a document are defined to be contiguous blocks of text (typically spanning several paragraphs), roughly 'about the same thing', with segment boundaries indicative of topic shifts, and/or changes in themes of discussion. Segmentation is informed by analysis of lexical repetition; in its turn, it supplies additional data concerning topicality, which now becomes available to the base summarizer sentence selection strategies.

In a sentence extraction-based model of summarization, a summary incorporating sentences from each segment will uniformly represent all sub-stories in a document. The notion here is to avoid indiscriminate joining, in the summary, of sentences which are far apart in the source document, or which belong to different topical chains. Moreover, by means of a mechanism which would pick the sentence(s) within a segment which are indicative of its main topic, we can ensure that the summary will retain 'traces' of *all the main topics* in the original document.

This is more than just an intuition. A failure analysis carried out during the development and training of our base summarizer (see 4.1 below) showed that 30.7% of a particular category of error could be prevented by a heuristic sensitive to the logical structure of documents, which would enforce that each (topical) section is represented in the summary. Additional 15.2% of failures could also be avoided if the summarizer was capable of detecting sub-stories within a single section, identifying leading/trailing noise (see 3 below), and so forth. Thus almost half of the errors (in a certain summarization regime, at least) could have been avoided by using a segmentation component.

This shows how a particular lexical cohesion factor (in this case simple repetition) can enhance an existing sentence selection strategy—even if such a strategy has been devised *without prior knowledge* of cohesion analysis. The specific approaches to being sensitive to foci of attention within a segment, and topic shifts between segments, may vary; as we will see (in Section 4), these will depend on other environment settings for the summarizer. Given the right operational environment, even very simple heuristics—such as, for instance, take the first sentence from each segment—have remarkably noticeable impact.

In essence, this paper argues that a lexical repetition-based model of linear segmentation offers highly

[2] As opposed to hierarchical; see (Marcu, 1997).

plausible schemes for deriving sentence-based summaries with certain discourse properties, as a result improving upon an already respectable system.

What follows is organized in four main sections. Our summarizer benefits from a number of linguistic analysis filters; these, as well as some additional details of the summarization and segmentation functions, are described next (Section 2). We focus in particular on how the higher level content analysis functions make use of lower level shallow linguistic processing, in order to obtain a richer model of the document(s) domain, and to leverage a cohesion metric for sub-story identification; the summarization and segmentation functions are thus described in some detail. Strategies for using segmentation results by the summarization component are discussed next (Section 3). We also briefly touch upon the issue of whether cohesion analysis is strictly a component for internal use by summarization, or whether it has other uses, of direct relevance and interest to the end user. Section 4 presents the results from a number of experiments, comparing the performance of summarization alone to segmentation-enhanced summarization. Our comparative evaluation uses the same methodology and testbed as originally developed for the summarizer training, and competitive evaluation; we therefore outline the evaluation testbed environment. Finally, we argue that it is possible, and beneficial, to make 'visual sense' of the notions we use: salience, topics, summary sentences, discourse segments, context, and so forth (Section 5). We conclude in Section 6 with an assessment of the overall utility of 'cheap' approximations to lexical cohesion measures, specifically from the point of view of enhancing a fully operational summarizer engine.

2 Technology base

The summarization system discussed here makes use of shallow linguistic functions. These are available 'for free, as the summarizer is an integral component of a significantly larger infrastructure for document processing and analysis, comprising a number of interconnected, and mutually enabling, linguistic filters. The infrastructure (hereafter referred to as TEXTRACT) is designed from the ground up to perform a variety of linguistic feature extraction functions. Some of these functions are relatively straightforward, e.g. single pass tokenisation, lexical look-up and morphological analysis. Others are complex, such as technical terminology extraction, and aggregation of salient phrasal units across large multi-document collections. To a large extent these characteristics of our document processing environment define the basic design decisions concerning the specifics of our summarizer: its sentence selection mechanisms utlilizes salience ranking of phrasal units in individual documents, viewed against a background of the distribution of phrasal vocabulary across a large multi-document collection.

2.1 Linguistic filters

At some level of generalization, TEXTRACT is a robust text analysis system that focuses on the identification of proper names and technical terms, since these are most likely to carry the bulk of the semantic load in a document. In addition to simple identification of certain phrasal types, however, the system is also capable of identifying their variants (contractions, abbreviations, colloquial uses, and so forth) in individual documents in a multi-document collection. This is the basis for a collection vocabulary of canonical forms and variants with statistical information concerning their distribution behaviour and prominence patterns across the collection. The collection vocabulary and statistics are used in the summarizer's salience calculation, which, in turn, is a significant component of the sentence-level score that selects the sentences for extraction (see 2.2).

Most of the linguistic analysis utilized by the summarizer is derived through a variety of shallow techniques. This is partly motivated by the requirements of an operational and robust system capable of efficient processing of thousands of documents/gigabytes of data. The TEXTRACT system, known commercially as INTELLIGENT MINER FOR TEXT, is an IBM product which has been successfully deployed

in a number of operational information management environments. At the same time, the system provides a testbed for studying the extent to which higher level semantic and discourse functions can be realized from a shallow linguistic base (Kennedy and Boguraev (1996) discuss this in some detail).

Fundamentally, our summarizer is a frequency-based system; thus it is ideally positioned to exploit TEXTRACT's functions for linguistic analysis, filtering, and normalization. Morphological processing allows us to link multiple variants of the same word, by normalizing to lemma forms. Proper name identification (Wacholder, Ravin, and Choi, 1997) is enhanced with context disambiguation, named entity typing, and variant normalization; as a result the system's frequency analysis is more precise, and less sensitive to noise; ultimately, this leads to more robust salience calculation. Normalization of different variants of the same concept to a canonical form is further facilitated by processes of abbreviations unscrambling, resolution of definite noun phrase anaphora, and aggregation across the entire document collection. The set of potentially salient phrases is enriched by the identification and extraction of technical terms (Justeson and Katz, 1995); this enables the recognition of certain multi-word concepts mentioned in the document, with discourse properties indicative of high topicality value, which is also directly relevant to salience determination.

Each document in a multi-document collection is analyzed individually. All 'content' words (non-stop words, in information retrieval terminology), as well as all the phrasal units identified by the TEXTRACT linguistic filters, are deemed to be *vocabulary items*, indexed via their canonical forms. With a view to future extensions of the base summarization function (see 6), these retain complete contextual information about the variants in which they have been encountered, as well as the local context of each occurrence. The vocabulary items are counted and aggregated across documents to form the *collection vocabulary*. In addition to all the canonical forms and variants detected during collection processing, the collection vocabulary contains the *composite frequency* of each canonical form, and its *information quotient*, a statistical measure of the distribution of a vocabulary item in the collection. Aggregating together similar items from different documents (cross-document co-reference) is far from straightforward for multi-word items; however, being able to carry out a process of cross-document coreference resolution is clearly a further enabling capability for obtaining more precise collection statistics (Kazi and Ravin, 2000). A general pronominal anaphora resolution capability (Kennedy and Boguraev, 1996) not only contributes additionally to the quality of the collection statistics, but also facilitates analysis of lexical cohesion, to the extent that one of the cohesion factors at play in large texts is directly related to anaphora.

In addition to the domain vocabulary, the summarizer also has access to the *document structure*. TEXTRACT incorporates a document structure builder, which constructs a hierarchical representation of the document. Content and layout metadata are separated. A document structure tree explicitly encodes information concerning document structural components: appearance and layout tags; document title; abstract, and other front matter; section, subsection, etc. headings; paragraphs, themselves composed of sentences; tables, figures, captions, and other 'floating' objects; side-bars and other kinds of text extraneous to the main document narrative; and so forth. At present, document structure is constructed by 'shadowing' markup (HTML, GML, etc.) parsing. For documents which lack markup tags, structure determination is carried out on the basis of two-dimensional (page) layout cues. Additional discourse-level annotations may also be recorded in the document structure: examples are cue phrases marking rhetorical relations, quoted speech, and so forth. Such elements contribute directly to the summarizer's set of heuristics, as well as inform the discourse segmentation process.

2.2 Salience-driven summarization

As an integral part of TEXTRACT's document processing environment, the summarizer makes extensive use of the output of TEXTRACT's linguistic filters. Fundamentally a frequency-based system, it also

exploits linguistic dimensions beyond single word analysis, not unlike the approach argued for by Aone et al. (1997). The depth of linguistic analysis and the degree to which the filters inform each other account for the summarizer's access to a richer source of domain knowledge than most other frequency-based systems.

Frequency alone is a poor indicator of term salience, even when the stop words are ignored. Unlike early frequency-based techniques for sentence selection, we utilize the more indicative *inverse document frequency* measure, adapted from information retrieval (Salton and McGill, 1983; Brandow, Mitze, and Rau, 1995), in which the relative frequency of an item in the document is compared with its relative frequency in a background collection. While facilitating more precise term salience, this is, unfortunately, at the cost of requiring for the summarizer's operation access to background collection statistics; we return to this issue below.

Sentence selection is driven by the notion of salience; the summary is constructed by identifying and extracting the most salient sentences in the source document. The *salience score of a sentence* is derived partly from the salience of the vocabulary items (including single-token words, multi-word names, abbreviations, and multi-word terms) in it, and partly from its position in the document structure (e.g. section-initial, paragraph-internal, and so forth) and the salience of the surrounding sentences. The vocabulary items are looked up in the collection vocabulary database by a statistical component that calculates, for each one, its inverse document frequency. The calculation compares the relative frequency of each item t in the document with the relative frequency of the item in the collection. We define the item's *salience score* to be this inverse document frequency measure (in the formula below, N_{Coll} and N_{Doc} refer to, respectively, to the number of items in the collection, and document).

$$Salience(t) = \log_2 \frac{N_{Coll}/freq(t)_{Coll}}{N_{Doc}/freq(t)_{Doc}}$$

Salient items are items occurring more than once in the document, whose salience score is above an experimentally determined cutoff, or items appearing in a strategic position in the document structure (e.g. title, headings, etc.). All others are assigned zero salience. The score for a sentence is made up of two components. The *salience* component is the sum of the salience scores of the items in the sentence. The *structure* component reflects the sentence's proximity to the beginning of the enclosing paragraph, and the paragraph's proximity to the beginning and/or end of the document. Structure score is secondary to salience score; sentences with no salient items get no structure score.

A set of heuristics are in place to address some of the coherence-related problems discussed in Section 1. For example, under certain conditions, a sentence might be selected for inclusion in the summary, even if it has low, or zero, score: sentences that immediately precede higher scoring ones in a paragraph may get promoted by virtue of an 'agglomeration rule', the operation of which is controllable from the client interface. Agglomeration is an inexpensive way of preventing dangling anaphors without having to resolve them. Another problem for sentence-based summarizers, also discussed in Section 1, is that of thematic under-representation (or, loosely speaking, coverage). This is addressed by another rule, the 'empty section' rule, which is of particular interest for this paper. Longer documents may be unevenly represented in a sentence-extracted summary. At least for documents with multiple sections marked with headings, or news digests containing multiple stories, an 'empty section' rule aims to ensure that each section/story is represented in the summary by forcing inclusion of its highest scoring sentences (if all sentence scores are zero, the first sentence of the section is used for the summary). In general, the sentence selection process operates under certain constraints. For example, sentences are not incorporated in a summary if they are too short (five words or less) or if they contain direct quotes (more than a minimum number of words enclosed in quotation marks).

We have described here a general purpose summarization capability, which makes extensive use both

of small scale linguistic information (term phrasal patterns) and large scale statistical information (term distribution patterns). As established in the NIST-sponsored SUMMAC evaluation (Hand and Sundheim, 1998), it turns out that such a system compares well with others, differently designed, ones. Still, with the exception of the heuristic rules outlined earlier in this section, the summarizer is operating without any focused analysis of cohesion factors in the input text. Hence the primary departure point for this work, namely: can the performance be improved, if we take into account lexical cohesion in the source?

Two additional questions arise when we consider some design characteristics of the summarizer: it is somewhat genre dependent (it performs best on news stories or news feature articles), and it crucially relies on a database of background statistics. At issue, then, are situations where the input documents are longer than the average length of a news story, as well as when no background collection exists for them. Neither of these situations is extraordinary. It is easy to conceive of document collections in a different genre: scientific articles, patent descriptions, financial reports, and so forth, all exhibit length significantly beyond what the current summarizer is designed to represent. Furthermore, new documents are created all the time; by definition, these do not belong to any background collection. It may take time to accumulate such a collection and analyze it; it may be impractical to store the vocabulary statistics of such a collection; it may be the case that existing collections do not adequately reflect the domain and genre of new documents.

We have chosen to address all of these questions by making the summarizer aware of certain discourse-level features of the document, and in particular, by leveraging the topic shifts in it; to this end, the TEXTRACT infrastructure has been augmented with a function for linear discourse segmentation.

2.3 Linear discourse segmentation

Segmentation is a document analysis function which directly exploits one of the core text cohesion factors, patterns of *lexical repetition* (1.1), for identifying some baseline data concerning the distribution of topics in a text. In particular, discourse segmentation is driven by the determination of points in the narrative where perceptible discontinuities in the text cohesion are detected. Such discontinuities are indicative of topic shifts. Following the original idea of *lexical chains* (Morris and Hirst, 1991), subsequently developed specifically for the purposes of segmentation of expository text (Hearst, 1994), we have adapted an algorithm for discourse segmentation to our document processing environment. In particular, while remaining sensitive to the distribution of "terms" across the document, and calculating similarity between adjacent text blocks by a cosine measure, our procedure differs from that in (Hearst, 1994) in several ways.

- ❏ We only take into account content words (as opposed to all terms yielded by a tokenization step);
- ❏ These are normalized to lemma forms;
- ❏ 'Termhood' is additionally refined to account for multi-word sequences (proper names, technical terms, and so forth, as discussed in 2.1), as well as for some (limited) notion of co-reference, where different name variants get aggregated into the same canonical form;
- ❏ The cohesion calculation function is biased towards different types of possible break points: thus certain cue phrases (*"However"*, *"On the other hand"*) unambiguously signal a topic shift; document structure elements—such as sentence beginnings, paragraph openers, and section heads (2.1)—are exploited for their 'pre-disposition' to act as likely segment boundaries; and so forth;
- ❏ The function is also adjusted to reduce the noise from block comparisons where the block boundary—and thus a potential topic shift—falls at unnatural break points (such as the middle of a sentence).

By making segmentation another component within the TEXTRACT's document processing environment, we are able to utilize, transparently, the results of processes such as *lexical and morphological lookup*,

document structure identification, and *cue phrase detection*. Likewise, the segmentation's output is naturally incorporated in an annotation superstructure which records the various levels of document analysis: discourse segments are just another type of a 'span' (annotation) over a number of sentences, logically akin to a paragraph (Bird and Liberman, 1999).

Apart from the adjustments and modifications outlined above, we use essentially the same formula as Hearst's for computing lexical similarity between adjacent blocks of text b_1 and b_2 (t denotes a discourse element term identified as such by TEXTRACT's prior processing, ranging over the text span of the currently analyzed block; ω_{t,b_N} is the normalized frequency of occurrence of the term in block b_N):

$$sim(b1, b2) = \frac{\Sigma_t \omega_{t,b_1} \omega_{t,b_2}}{\sqrt{\Sigma_t \omega_{t,b_1}^2 \Sigma_t \omega_{t,b_2}^2}}$$

Figure 1 illustrates the results of the 'raw' segmentation process, as applied to a standard length (approximately two pages) news article.

```
[snt  2] at: 0013--0043 [P] "Throughout the 1... " penalty: ( 77.393265)                      ..................  [  ]
[snt  3] at: 0044--0084 [P] "During their nin... " penalty: ( 77.491781)                      ..................  [  ]
[snt  4] at: 0085--0102     "The toughest Sov... " penalty: ( 98.453205)              ..............................  [  ]
[snt  5] at: 0103--0117 [P] "But neither carp... " penalty: ( 79.945415) [S]                +-------------------  [ 5]
[snt  6] at: 0118--0126     "Afghanistan has ... " penalty: (101.610705)            ..............................  [  ]
[snt  7] at: 0127--0173     "Its terrain favo... " penalty: ( 98.562303)            ..............................  [  ]
[snt  8] at: 0174--0202     "It is uncertain ... " penalty: ( 92.657816)               ...........................  [  ]
[snt  9] at: 0203--0260 [P] "The camps, hidde... " penalty: ( 66.978529) [S]                        +------------  [ 9]
[snt 10] at: 0261--0284 [P] "The Afghan resis... " penalty: ( 70.478605)                         ..............  [  ]
[snt 11] at: 0285--0332     "And the territor... " penalty: ( 87.576509)                     ..................  [  ]
[snt 12] at: 0333--0354 [P] "The CIA's milita... " penalty: ( 71.787064) [S]                    +--------------  [12]
[snt 13] at: 0355--0378     "And some of the ... " penalty: ( 93.124730)                ............................  [  ]
[snt 14] at: 0379--0410 [P] "From those same ... " penalty: ( 73.534306)                       ................  [  ]
[snt 15] at: 0411--0421 [P] "Thousands of muj... " penalty: ( 75.611965)                       ................  [  ]
[snt 16] at: 0422--0446     "Soviet accounts ... " penalty: ( 97.129682)              ..............................  [  ]
[snt 17] at: 0447--0476     "After a decade o... " penalty: ( 93.898393)               ...........................  [  ]
[snt 18] at: 0477--0511 [P] ""This was the mo... " penalty: ( 57.492338) [S]                           +--------  [18]
[snt 19] at: 0512--0552 [P] "U.S. officials s... " penalty: ( 59.910953)                           ..........  [  ]
[snt 20] at: 0553--0568     "They said the da... " penalty: ( 88.244649)                  ...................  [  ]
[snt 21] at: 0569--0600 [P] "But the communic... " penalty: ( 59.745026) [S]                         +---------  [21]
[snt 22] at: 0601--0624     "The strongest po... " penalty: ( 80.532043)                     ..................  [  ]
[snt 23] at: 0625--0641 [P] "And history does... " penalty: ( 65.539264)                          ............  [  ]
[snt 24] at: 0642--0669 [P] "Bin Laden has sa... " penalty: ( 61.233445) [S]                          +----------  [24]
[snt 25] at: 0670--0697     "He was most stro... " penalty: ( 73.174999)                        ................  [  ]
[snt 26] at: 0698--0736     "After the fall o... " penalty: ( 84.207023)                   ....................  [  ]
[snt 27] at: 0737--0765 [P] "The more militan... " penalty: ( 67.018487) [S]                     +------------  [27]
[snt 28] at: 0766--0794     "A year after the... " penalty: ( 91.185994)                ............................  [  ]
[snt 29] at: 0795--0819 [P] "No amount of mon... " penalty: ( 71.567491)                        ...............  [  ]
[snt 30] at: 0820--0849     "The chaos that t... " penalty: ( 91.413237)                ............................  [  ]
[snt 31] at: 0850--0890 [P] "In the nine year... " penalty: ( 61.808809) [S]                          +----------  [31]
[snt 32] at: 0891--0922     "In those years, ... " penalty: ( 88.473182)                  ...................  [  ]
[snt 33] at: 0923--0930     "Bin Laden sponso... " penalty: ( 83.881455)                    ..................  [  ]
[snt 34] at: 0931--0960 [P] "In a 1994 interv... " penalty: ( 64.709675) [S]                       +-----------  [34]
[snt 35] at: 0961--0971 [P] ""There are many ... " penalty: ( 78.130276)                      ..................  [  ]
[snt 36] at: 0972--0989     ""We have had Egy... " penalty: (103.308307)            ..............................  [  ]
[snt 37] at: 0990--1011     "U.S. officials s... " penalty: (100.272182)             ............................  [  ]
[snt 38] at: 1012--1053 [P] "Bin Laden, strip... " penalty: ( 70.019151) [S]                    +--------------  [38]
[snt 39] at: 1054--1113 [P] "He said in an in... " penalty: ( 72.609082)                        ...............  [  ]
[snt 40] at: 1114--1170 [P] "It is unclear wh... " penalty: ( 68.709284)                         ..............  [  ]
```

Figure 1: 'Raw' discourse segmentation: cohesion curve

Local minima (the "+" points) in the shape of the cohesion curve (outlined on the right) denote low penalty for positing a segment boundary at that point; generally speaking, these are the points where topic shifts occur. Segmentation results are 'overlayed' onto the document structure (this is marked up in the printout by sentence offset points "`[snt]`", paragraph "`[P]`" marks, and segment boundary "`[S]`" breaks).

The topic shifts are illustrated, schematically, in Figure 2 (only half of the document is shown).

Throughout the 1980s, the Soviet Union threw almost every weapon it had, short of nuclear bombs, at the Afghan camps attacked by the United States last week.

During their nine-year occupation of Afghanistan, the Soviets attacked the camps outside the town of Khost with Scud missiles, 500-pound bombs dropped from jets, barrages of artillery, flights of helicopter gunships and their crack special forces. The toughest Soviet commander in Afghanistan, Lt. Gen. Boris Gromov, personally led the last assault.

But neither carpet bombing nor commandos drove the Afghan holy warriors from the mountains. Afghanistan has a long history of repelling superpowers. Its terrain favors defenders as well as any in the world, whether their opponents, like the Soviets, are trying to defeat them on the ground or whether, like the United States, they are trying to disperse, deter and disrupt them. It is uncertain that the United States, which fired dozens of million-dollar cruise missiles at those same camps on Thursday, can do better than the Soviets.

The camps, hidden in the steep mountains and mile-deep valleys of Paktia province, were the place where all seven ranking Afghan resistance leaders maintained underground headquarters, mountain redoubts and clandestine weapons stocks during their bitter and ultimately successful war against Soviet troops from December 1979 to February 1989, according to American intelligence veterans.

The Afghan resistance was backed by the intelligence services of the United States and Saudi Arabia with nearly $6 billion worth of weapons. And the territory targeted last week, a set of six encampments around Khost, where the Saudi exile Osama bin Laden has financed a kind of "terrorist university," in the words of a senior U.S. intelligence official, is well known to the CIA.

The CIA's military and financial support for the Afghan rebels indirectly helped build the camps that the United States attacked. And some of the same warriors who fought the Soviets with the CIA's help are now fighting under bin Laden's banner.

From those same camps, the Afghan rebels, known as mujahedeen, or holy warriors, kept up a decadelong siege on the Soviet-supported garrison town of Khost.

Thousands of mujahedeen were dug into the mountains around Khost. Soviet accounts of the siege of Khost during 1988 referred to the rebel camps as "the last word in NATO engineering techniques." After a decade of fighting during which each side claimed to have killed thousands of the enemy, the Afghan rebels poured out of their encampments and took Khost.

"This was the most fiercely contested piece of real estate in the 10-year Afghan war," said Milt Bearden, who ran the CIA's side of the war from 1986 to 1989.

U.S. officials said their attack was intended to deter bin Laden, whom they call the financier and intellectual author of this month's bombings of two American embassies in Africa, which killed 263 people, including 12 Americans. They said the damage inflicted on the Khost camps was "moderate to heavy."

Figure 2: 'Raw' discourse segmentation: topic shifts

Most applications of segmentation to date, typically in information retrieval, are concerned with the identification of segment boundaries, with a view of offering a more refined 'positioning' of a query answer in its document context (Hearst, 1995; Salton et al., 1996; Ponte and Croft, 1997; Beeferman, Berger, and Lafferty, 1997; Reynar, 1998). In contrast, we are primarily interested in leveraging the content of the segments, because it is the items at the focus of attention we want to identify and retain for inclusion in the summary. In addition, we use the segmentation results (together with the name and term identification and salience calculation) in order to make sure that all the base data for inferring the topical make-up of the document is available to the user.

This raises two related questions. The first concerns the relationship between segmentation and summarization: is segmentation a strictly "under the covers", service function used by the summarizer, or might the results of discourse segmentation be of any interest, and use, to the end user? We discuss, in the following section, strategies for incorporating segmentation results in the summary generation process. However, unlike (Kan, Klavans, and McKeown, 1998) whose work also seeks to leverage linear segmentation for the explicit purposes of document summarization, we further take the view that with an appropriate interface metaphor—where the user has an overview of the relationships between a summary sentence, the key salient phrases within it, and its enclosing discourse segment—a sequence of visually demarkated segments can impart a lot of information directly leading to the in-depth perception of the summary, as it relates to the full document. The second question thus concerns the features of such an interface. We return to this point below (see 5).

3 Using segmentation in summaries

Common intuitions suggest a number of strategies for enhancing summarization by incorporating results of linear discourse segmentation. As points of topical shifts in the text are 'published' into the document structure (see 2.3), the summarizer transparently becomes aware of the segmentation results. We also make arrangements for a mechanism whereby certain strategies for incorporating segmentation results into the summarization process are easy to cast in summarizer terms.

Thus, for instance, a heuristic which would require that each segment is represented in the summary can be naturally expressed by treating segments as sections, and strictly enforcing the 'empty section' rule (see 2.2). A strategy which requires the selection of a segment-initial sentence for the summary can be enabled simply by boosting the salience score for that sentence above a known threshold. A decision to drop an anecdotal (or otherwise peripheral; see below) segment from consideration in summary generation would be realised by setting, as a last step prior to summary generation, the sentence salience scores for all sentences in the segment to zeros.

Such strategies are discussed in more detail in the next section, as they naturally belong with their evaluation. Here, we highlight a few observations concerning the overall benefits that segmentation brings to summarization. Thus, in addition to facilitating various schemes for deriving sentence-based summaries with certain discourse and rhetorical properties, it turns out that at least one such scheme also allows the summarizer to operate without a need for background corpus statistics. Ideally, availability of a backgrouhd collection should not always be assumed, because it limits the applicability of otherwise a general purpose summarizer (see 2.2). It may also be impractical, for a variety of reasons (no time for processing, sparse storage resources, rapid incremental growth, and so forth), to collect statistics, even if sufficiently large and representative data sample was available.

Given the underpinnings of discourse segmentation, a reasonable strategy for purely segmentation-informed summaries is to include in them the first sentence from each segment. This is a better, and cheaper, solution than the highly genre-dependent approach of supplying a 'generic' background collection, against which summaries could be generated even for documents which are not *a priori* part of the collection. This not only offers an operational bonus for the summarizer, but in certain cases improves upon the quality of frequency-informed summaries (see 4 below).

Other uses for a segmentation component in summarization context optimize the information content of the source (e.g. by selecting input fragments which contain relevant information), and maximize its reuse. Occasionally, the document contains 'noise'—this may be in the form of *anecdotal leads*, *closing remarks* tangential to the main points of the story, *side-bars*, and so forth—which should not be considered as source for summary sentences. Linear segmentation sensitive to topic shifts and document structure would identify such source fragments and remove them from consideration by the summarizer. Conversely, in certain genres of news reporting a contiguous document fragment (typically towards the beginning or the end of the document) functions as a *precis* of the story: having it identified as a segment in a certain discourse position is highly relevant to the summarization task.

We also use segmentation to handle long documents more effectively. While the collection-based salience determination works reasonably well for the average-length news story, it has some disadvantages. For longer documents, with requisite longer summaries, the notion of salience degenerates, and the summary becomes just an incoherent collection of sentences. (Even if paragraphs, rather than sentences, are used to compose the summary, as e.g. Mitra, Singhal, and Buckley (1997) do, the same problems of coherence degradation and topical under-representation, remain.) We use segmentation to identify contiguous sub-stories in long documents, which are then individually passed on to the summarizer; the results of sub-story summaries are 'glued' together.

4 Evaluation

The impact of various strategies upon summarizer output quality was assessed by using as baseline an evaluation corpus of full-length articles and their 'digests', from *The New York Times*. There are advantages, and disadvantages, to this approach. Setting aside the issue of whether the effect of one technology on another is best highlighted by means of a task-based evaluation (see 4.1 below), such a decision ties us to a particular set of data. On the positive side, this offers a realistic baseline against which to compare strategies and heuristics; on the negative side, if a certain type of data is missing from the evaluation corpus, there is little hard evidence for judging the effects of strategies and heuristics on such data.

In the remainder of this section we describe our evaluation environment, and then focus on presenting the results for small-to-average size documents (the collection comprises just over 800 texts, less than half of which are over 10K, and virtually none are over 20K; the byte count includes HTML markup tags; in terms of number of sentences per document, very few of these longer documents are over 100 sentences long).

4.1 Summarization evaluation testbed

Evaluating summarization results is not trivial, at least because there is no such thing as the best, or 'canonical', summary—especially when the summary is constructed as an extract. The purposes of such extracts vary; so do human extractors. Sentence extraction systems may be evaluated by comparing the extract with sentences selected by human subjects (Edmundson, 1969). This is a (superficial) objective measure that clearly ignores the possibility of multiple right answers. Another objective measure compares summaries with pre-existing abstracts using a suitable method for mapping a sentence in the abstract to its counterpart in the document (Kupiec, Pedersen, and Chen, 1995). Subjective measures, even though still less satisfying, can also be devised: for instance, summary acceptability has been proposed as one such measure (Brandow, Mitze, and Rau, 1995). Other evaluation protocols share the primary feature of being *task-based*, even though details may vary. Thus performance may be measured by comparing browsing and search time as summary abstracts and full-length originals are being used (Miike et al., 1994); other measures look at recall and precision in document retrieval (Brandow, Mitze, and Rau, 1995); or recall, precision, and time required in document categorization, namely assessing whether a document has been correctly judged to be relevant or not, on the basis of its summary alone (Mani, Firmin, and Sundheim, 1999).

We built our own environment for baseline summarizer evaluation, as part of its development/training cycle. This same environment was used in analyzing the impact of discourse segmentation on the summarizer's performance. Vocabulary statistics for a background collection were gathered from analyzing 2334 *New York Times* news stories. Sentences in digests for 808 news stories and feature articles were automatically matched with their corresponding sentences in the full-length documents, using a vector-based program for text similarity analysis (Prager, 1999) that was able to map source to digest sentences even when moderate differences existed between the two. Digests range in length from 1 to 4 sentences. Since we were particularly interested in longer stories, as well as stories in which the first sentence in the document did not appear in the digest, their representation in the test set, 38%, is larger than their distribution in the newspaper.

Given that digests are inherently short, this evaluation strategy is somewhat limited in its capability of fully assessing segmentation effects on summarization of long documents. Nonetheless, a number of comparative analyses can be carried out against this baseline collection, which are indicative of the interplay of the various control options, environment settings, and TEXTRACT filters used. One parameter, in particular, is quite instrumental in tuning the summarizer's performance, to a large extent because it

is directly related to length of the original document: size of the summary, expressed either as number of sentences, or as percentage of the full length of the original. In addition to a clear intuition (namely that the size of the summary ought to be related to the size of the original), varying the length of the summary offers both the ability to measure the summarizer's performance against baseline summaries (i.e. our collection of digests), and the potential of dynamically adjusting the derived summary size to optimally represent the full document content, depending on the size of that document.

Our experiments vary the granularity of summary size. In principle, the performance of a system which does absolute sentence ranking, and systematically picks the N 'best' sentences for the summary, should not depend on the summary size. In our case, the additional heuristics for improving the coherence, readability, and representativeness of the summary (see 2.2) introduce variations in overall summary quality, depending on the compaction factor applied to the original document size.

A representative spectrum for the test corpus we use is given by data points at:

- ❏ *digest size* (i.e. summary exactly the size, expressed as number of sentences, of the digest);
- ❏ summary of precisely *4 sentences*;
- ❏ summary at *10% of the size* of the full length document; and
- ❏ summary at *20% of the document*.

Not surprisingly (for a salience-based system), the summarization function alone, without discourse segmentation, benefits from larger summary size. Although the recall rate is higher still for longer summaries, it is not a measure of the overall quality of the summary because of the inherently short length of the digest.

4.2 Segmentation effects on summarization

Our experiments compare the base summarization procedure, which calculates object salience with respect to a background document collection (see 2.2), with enhanced procedures incorporating several different strategies for using the notions of discourse segments and topic shifts.

These elaborate the intuitions underlying our approach to leveraging lexical cohesion effects (see 1.2). The experiments fall in either of two categories. In an environment where a background collection, and statistics, cannot be assumed, a summarization procedure was defined to take selected (typically initial) sentences from each segment; this appeals to the intuition that segment-initial sentences would be good topic indicators for their respective segments. The other category of experiment focused on enriching the base summarization procedure with a sentence selection mechanism which is informed by segment boundary identification and topic shift detection.

In combining different sentence selection mechanisms, several variables need adjustment to account for relative contributions of the different document analysis methods, especially where summaries can be specified to be of different lengths. Given the additional sentence selection factors interacting with absolute sentence ranking, we again set the granularity of summary size at three discrete steps, mirroring the evaluation of the original summarizer: summaries can be requested to be precisely 4 sentences long, or to reflect source compaction factor of 10% or 20% (see 4.1).

We experimented with two broad strategies for actively incorporating topical information into the summary. One approach of explicitly using 'topic openers' adds segment-initial sentences to the set of sentences already selected by the salience calculation mechanism. The notion was to exert finer control over the number of sentences selected via salience (N), and 'pad' the summary to its requested size with sentences selected from segments by invoking the 'empty segment' rule (aka 'empty section' rule, see 2.2). Special provisions were made to account for the fact that segmentation would naturally always select the first sentence in the document.

Table 1 shows the rates of recall, against the digests, for the three major summarization regimes defined by different summary granularities. In the table, the label "SUM" denotes a regime running only the base summarization component alone: frequency-based, salience-informed procedure, as described in 2.2. "SEG" represents summarization by segmentation alone: summaries in this regime are made up of segment-initial sentences. We use the label "SUM+SEG" to denote a strategy for constructing summaries by combining sentences with high salience (as calculated by SUM method) with sentences with discourse prominence (identified as segment-initial by SEG).

In our experiments, N was set at 2, 4, and 8; the results in the table are shown at $N = 4$. It is reasonable to devise a strategy which dynamically will set N depending on the requested size of summaries, for instance, $N = 2$ for smaller summaries (at 4 sentences long), and $N = 4$ or $N = 8$ for summaries at 10% and 20% respectively.

Since segmentation effects are clearly very different across different sizes of source document, our experiments were additionally conducted at sampling the document collection at different sizes of the originals: the corpus was split into four sections, grouping together documents less than 7.5K characters long, 7.5–10K, 10–19K, and over 19K; for brevity, the table (1.a) encapsulates a 'composite' result (denoted by the label *"All documents"*).

	4 sents	10%	20%
a.: All documents			
SEG	54.74	54.74	56.09
SUM	46.85	49.71	66.47
SUM+SEG	52.49	57.25	61.57
b.: All documents with > 1 digest sentence			
SEG	45.13	45.13	46.78
SUM	36.34	39.84	58.66
SUM+SEG	41.64	46.75	51.65
c.: All documents whose 1st sentence not in target digest			
SEG	31.12	32.73	33.99
SUM	29.93	39.96	61.71
SUM+SEG	32.53	41.45	47.96

Table 1: Summary data for segmentation effects on summarization

To get a better sense for the effects of different strategy mixes, we also show results for the same summarization regimes, on subsets of the test corpus; these were specifically constructed by removing documents whose digests are such that a default strategy would be guaranteed to pick a representative sentence. The intent was to observe performance on more challenging data. Thus 1.b, *"All documents with > 1 digest sentence"*, represents documents whose digests are longer than a single sentence; 1.c, *"All documents whose 1st sent is not in target digest"*, extracts a document set for which a baseline strategy automatically picking a representative sentence for inclusion in the summary would be inappropriate. These subset selection criteria explain the deterioration of overall results; however, what is more interesting to observe in the table is the relative performance of the three summarization regimes.

Further experiments investigated a different strategy for incorporating 'topic openers' in the summary. A bonus 'weight' was added to segment-initial sentences, to make them look like salient ones (in base summarizer terms), in an attempt to streamline a sentence selection method which does not rely so heavily on an 'empty segment' rule. This did not yield conclusive results.

Overall, leveraging some of the segmentation analysis is positively beneficial to summarization; the effects are particularly strong where short summaries are required. In addition, the summarization procedure defined to work from segmentation data alone shows recall rates comparable to, and in certain

situations even higher than, the original TEXTRACT function: this suggests that such a procedure is certainly usable in situations where background collection-based salience calculation is impossible, or impractical.

Finally, we emphasise a note of particular interest here: the complete set of data from these experiments makes it possible, for any given document, to select dynamically the summarization strategy appropriate to its size, in order to get an optimal summary for it, in any given information compaction regime.

5 Segment visualization and summary coherence

We take the view that segmentation is not only a subsidiary function for enhancing the quality of summarization, but a process which is of independent utility for the end user, as long as its results are integrated within an appropriate interface. There are two components to fully understanding a summary. First, there is the ability to see the span of a segment, together with some indication of its core topic stamps. This should be then complemented by the ability to use the summary as the primary navigation tool for mediating between it, and the full text of the original. Boguraev, Bellamy, and Kennedy (1999) discuss in some detail the first of these two issues, topic stamping of individual segments, and argue for a presentation metaphor where topics stamps are the primary navigation element into the full document content.

The second issue above is even more critical, since the only way of making some sense of the summary as a characterization of the full document is by being able to 'undo' the effects of ellision of material between each two adjacent summary sentences. However, in the canonical summarization framework, representating the fragments missing from the source is very hard to arrange for by means of a visual abstraction. This is a direct consequence of the problem of under-representation (see 1), because the client typically has no control over the extent of the material which falls below the sentence salience threshold.

Discourse segmentation, on the other hand, is intended to address this problem. Thus it also turns out to offer the means of a richer visual abstraction, which directly incorporates the notion of topic shifts at the interface.

Figure 3 presents a screen snapshot of a prototype front end to a segmentation-enhanced summarizer, which is capable of contextualising summary sentences, indicating the span of omitted material between them, and suggesting grouping of summary fragments to show topic highlighting. A crucial feature of this interface is that the two different information panes, the summary one on the left and the full length document on the right, are *synchronously scrollable*; furthermore, both displays are 'anchored' to the segment span visual abstraction—the vertical bar in the middle—which is the primary organizational device mediating both a view of the summary and a representation of topical content (both topic stamps and topic shifts).

Without discourse segmentation, this kind of visual metaphor would be very hard to render on a summary stream which does not have topical information in it: consider the very common practice of hyperlinking summary sentences to their counterparts in the full document source (Mahesh, 1997). Due to the under-representation problem, the summary (left) pane might be too sparse; visually, this would translate into mis-cueing the user whether what is seen in the summary pane is a complete summary, or a fragment whose continuation is only reachable after (arbitrary amount of) scrolling. Additional problems arise from lack of any data to facilitate the user in identifying topics missing from the summary in what would be a long passage in the right pane, without any topical (or other) annotation.

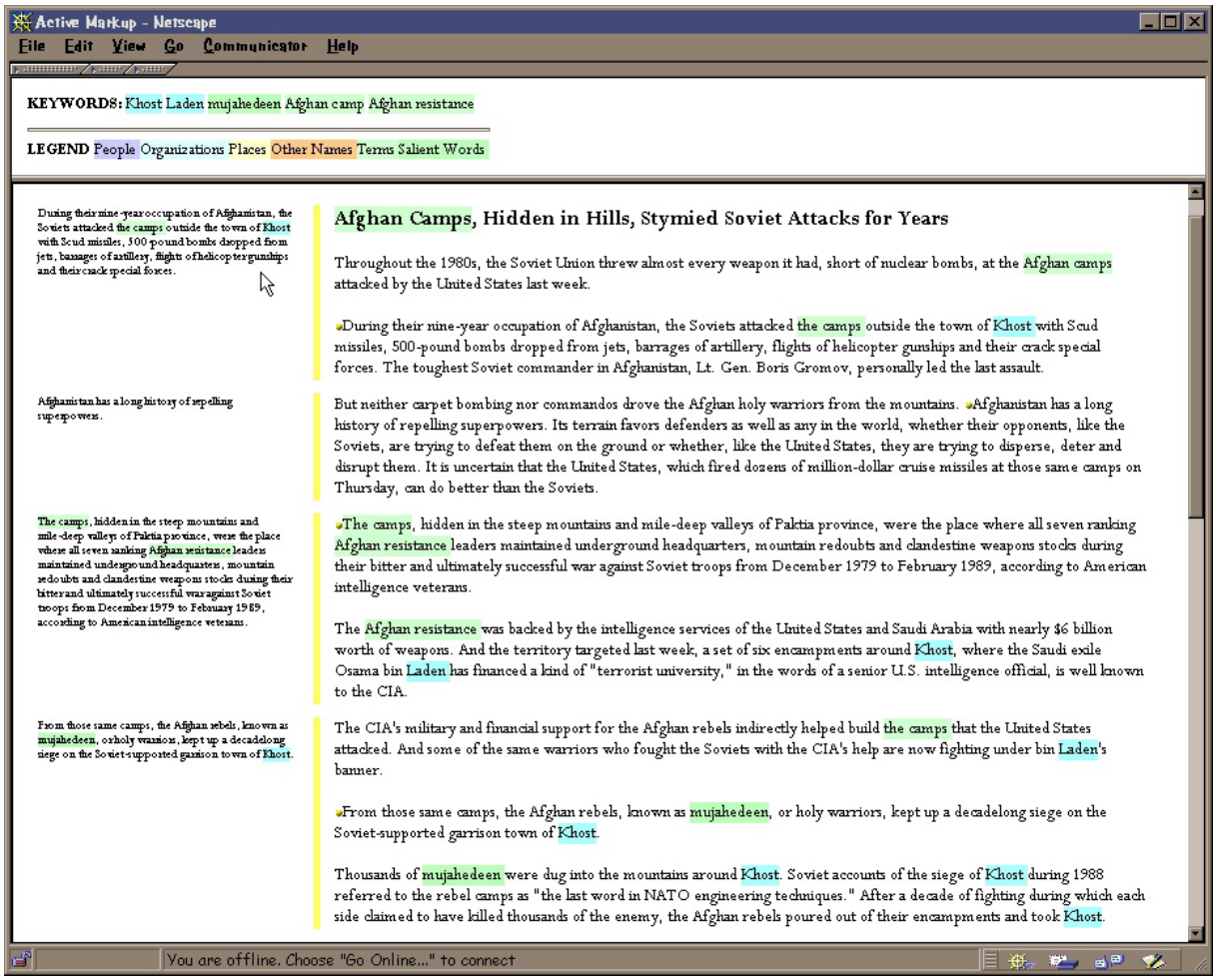

Figure 3: TEXTRACT summarizer: segmentation overlay

The interface makes use of additional features, some of them in common use in document content visualization environments: a hyperlink device facilitating attention switches between the summary and document panes, while retaining focus on a topical sentence; colour coding for marking and displaying salient vocabulary items; hot spots highlighting recurring occurrences of a salient item. We will not discuss these in detail here, as they are not directly related to the integration of segmentation and summarization functions (but see (Neff and Cooper, 1999)).

6 Conclusion

Starting from a class of problems inherent to summarization-by-sentence-extraction technology, we have proposed a strategy for alleviating some of the particularly jarring end-user effects in the summaries, which are due to coherence degradation, readability deterioration, and topical under-representation. Our approach is to aim for more cohesive summaries, by leveraging the lexical cohesion factors in the source document texts. As an initial experiment, we have looked at one particular factor, lexical repetition, and have developed a framework for integrating a discourse segmentation component capable of detecting shifts in topic, with a linguistically-aware summarizer which utilizes notions of salience and dynamically-adjustable size of the resulting summaries. By analyzing cohesion indicators in the discourse, segmentation identifies points in the narrative where sub-stories alternate; the summarization function uses the resulting set of discourse segments to derive more complete, informative and faithful summaries than ones extracted solely on the basis of sentence salience (with respect to a background

document collection).

A comparative evaluation of summarization with, and without, segmentation analysis shows that under certain conditions, segmentation-enhanced summarization is better than the base segmentation technology utilized in TEXTRACT. Some of these conditions can be expressed as a function of the original document length, and the document-to-summary ratio; thus, of particular interest is the fact that optimal strategy for combining the two technologies can be selected 'on the fly', depending on the type of input to be summarized.

Furthemore, having access to a segmentation component makes it possible to alleviate a serious shortcoming of summarizers like ours, which crucially depend on the statistics of a background collection: in situations where background collection-based salience calculation is impossible, or impractical, it is realistic to deliver summaries—of comparable quality, yet considerably cheaper to generate—derived by access to discourse segmentation information alone.

The research reported here is part of a larger effort focused on leveraging elements of the discourse structure for a variety of content characterisation tasks. Overall, we aim to build an infrastructure for recognizing and using a broad range of cohesive devices in text. Document summarization is just one application in the larger space of document content management; our long term goal is to develop a framework where summarization and other applications would be enabled by a rich substrate of linguistic analysis of lexical cohesion.

References

Aone, Chinatsu, Mary Ellen Okurowski, James Gorlinsky, and Bjornar Larsen. 1997. A scalable summarization system using robust NLP. In *Intelligent Scalable Text Summarization, Proceedings of a Workshop Sponsored by the Association for Computational Linguistics*, pages 66–73.

Barzilay, Regina and Michael Elhadad. 1999. Using lexical chains for text summarization. In Inderjeet Mani and Mark T. Maybury, editors, *Advances in automatic text summarization*. MIT Press, Cambridge, MA, pages 111–121.

Beeferman, D., A. Berger, and J. Lafferty. 1997. Text segmentation using exponential models. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*.

Bird, Steven and Mark Liberman. 1999. Annotation graphs as a framework for multidimensional linguistic data analysis. In *Proceedings of a Workshop, "Towards Standards and Tools for Discourse Tagging", 37th Annual Meeting of the Association for Computational Linguistics*, pages 1–10, Baltimore, MD.

Boguraev, Branimir, Rachel Bellamy, and Christopher Kennedy. 1999. Dynamic presentations of phrasally-based document abstractions. In *Hawaii International Conference on System Sciences* (HICSS-32)*: Understanding Digital Documents*, Maui, Hawaii, January.

Boguraev, Branimir, Yin Yin Wong, Christopher Kennedy, Rachel Bellamy, Sascha Brawer, and Jason Swartz. 1998. Dynamic presentation of document content for rapid on-line browsing. In *AAAI Spring Symposium on Intelligent Text Summarization*, pages 118–128, Stanford, CA.

Brandow, R., K. Mitze, and L. Rau. 1995. Automatic condensation of electronic publications by sentence selection. *Information Processing and Management*, 31(5):675–685.

Edmundson, H.P. 1969. New methods in automatic abstracting. *Journal of the ACM*, 16(2):264–285.

Fellbaum, Christiana, editor. 1999. WORDNET*: an electronic lexical database and some of its applications*. Cambridge, MA: MIT Press.

Halliday, M.A.K. and R. Hasan. 1976. *Cohesion in English*. London: Longman.

Hand, Therese F. and Beth Sundheim, editors. 1998. TIPSTER/SUMMAC *Summarization Analysis; Tipster Phase* III *18-Month Meeting"*, NIST, Fairfax, Virginia. Defense Advanced Research Project Agency. Working papers from SUMMAC conference.

Hearst, Marti. 1994. Multi-paragraph segmentation of expository text. In *32nd Annual Meeting of the Association for Computational Linguistics*, Las Cruces, New Mexico.

Hearst, Marti A. 1995. Tilebars: Visualization of term distribution information in full text information access. In *ACM SIGCHI Conference on Human Factors in Computing Systems*, Denver, CO.

Hoey, Michael. 1991. *Patterns of lexis in text*. Oxford, UK: Oxford University Press.

Justeson, John S. and Slava M. Katz. 1995. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1(1):9–27.

Kan, Min-Yen, Judith L. Klavans, and Kathleen R. McKeown. 1998. Linear segmentation and segment significance. In Eugene Charniak, editor, *Proceedings of the Sixth Workshop on Very Large Corpora*, pages 197–205, Montreal, Canada, August. Sponsored by ACL and ACL's SIGDAT.

Kazi, Zunaid and Yael Ravin. 2000. Who's who? Identifying concepts and named entities across multiple documents. In *Hawaii International Conference on System Sciences (*HICSS-33*)*.

Keller, Andrew. 1994. Common topics and coherent situations: interpreting ellipsis in the context of discourse inference. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 50–57, Las Cruces, NM.

Kennedy, Christopher and Branimir Boguraev. 1996. Anaphora for everyone: Pronominal anaphora resolution without a parser. In *Proceedings of COLING-96 (16th International Conference on Computational Linguistics)*, Copenhagen, DK.

Kupiec, Julian, Jan Pedersen, and Francine Chen. 1995. A trainable document summarizer. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 68–73, Seattle, Washington.

Mahesh, Kavi. 1997. Hypertext summary extraction for fast document browsing. In *Proceedings of AAAI Spring Symposium on Natural Language Processing for the World Wide Web*, pages 95–104, Stanford, CA.

Mani, Inderjeet, Therese Firmin, and Beth Sundheim. 1999. The TIPSTER SUMMAC text summarization evaluation. In *Proceedings of the Ninth Conference of the European Chapter of the ACL*, pages 77–85, Bergen, Norway, June. Association for Computational Linguistics.

Mani, Inderjeet and Mark T. Maybury, editors. 1999. *Advances in automatic text summarization*. Cambridge, MA: MIT Press.

Marcu, Daniel. 1997. From discourse structures to text summaries. In *Proceedings of ACL'97 Workshop on Intelligent, Scalable Text Summarisation*, pages 82–88, Madrid, Spain.

Miike, Seije, Etsuo Itho, Kenji Ono, and Kazuo Sumita. 1994. A full text retrieval system with a dynamic abstract generation function. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 152–161.

Mitra, Mandar, Amit Singhal, and Chris Buckley. 1997. Automatic text summarisation by paragraph extraction. In Inderjeet Mani and Mark T. Maybury, editors, *Proceedings of a Workshop on Intelligent Scalable text Summarization*, pages 39–46, Madrid, Spain. Sponsored by the Association for Computational Linguistics.

Morris, Jane and Graeme Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17:21–48.

Neff, Mary S. and James W. Cooper. 1999. ASHRAM: active summarization and markup. In *Hawaii International Conference on System Sciences* (HICSS-32)*: Understanding Digital Documents*, Maui, Hawaii, January.

Phillips, M. 1985. *Aspects of text structure: an investigation of the lexical organization of text*. Amsterdam: North Holland.

Phillips, M. 1989. The lexical structure of text. Discourse Analysis Monographs 12, English Language Research, University of Birmingham.

Ponte, J.M. and W.B. Croft. 1997. Text segmentation by topic. In *Proceedings of the First European Conference on Research and Advanced Technology for Digital Libraries*, pages 120–129.

Prager, John. 1999. LINGUINI: language identification for multilingual documents. In *Hawaii International Conference on System Sciences* (HICSS-32)*: Understanding Digital Documents*, Maui, Hawaii,

January.

Reynar, Jeff. 1998. *Topic segmentation: algorithms and applications*. Ph.D. thesis, University of Pennsylvania, Department of Computer and Information Science.

Salton, Gerald and Michael McGill. 1983. *An Introduction to Modern Information Retrieval*. McGraw-Hill.

Salton, Gerald, Amit Singhal, Chris Buckley, and Mandar Mitra. 1996. Automatic text decomposition using text segments and text themes. In *Seventh ACM Conference on Hypertext*, Washington, D.C.

Wacholder, Nina, Yael Ravin, and Misook Choi. 1997. Disambiguation of names in text. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pages 202–208, Washington, D.C., March.

Winter, E.O. 1979. Replacement as a fundamental function of the sentence in context. *Forum Linguisticum*, 4(2):95–133.