The effects of analysing cohesion on document summarisation

Branimir K. Boguraev and Mary S. Neff

IBM T.J. Watson Research Center, P.O. Box 704, Yorktown Heights, NY 10598, USA

bkb,neff@watson.ibm.com

Abstract

We argue that in general, the analysis of lexical cohesion factors in a document can drive a summarizer, as well as enable other content characterization tasks. More narrowly, this paper focuses on how one particular cohesion factor—simple lexical repetition—can enhance an existing sentence extraction summarizer, by enabling strategies for overcoming some particularly jarring end-user effects in the summaries, typically due to coherence degradation, readability deterioration, and topical under-representation. Lexical repetition is instrumental to, among other things, the topical make-up of a text, and in our framework a lexical repetition-based model of discourse segmentation, capable of detecting topic shifts, is integrated with a linguistically-aware summarizer utilizing notions of salience and dynamically-adjustable summary size. We show that even by leveraging lexical repetition alone, summaries are of comparable, and under certain conditions better, quality than the ones delivered by a state-of-the-art summarizer. This is encouraging for a broad research platform focusing on the recognition and use of cohesive devices in text for a range of content characterisation and document management tasks.

1 Introduction

This paper addresses a particular class of problems inherent to summaries derived by sentence extraction, namely the related issues of *coherence degradation*, *readability deterioration*, and *topical under-representation*. Fundamentally, these problems arise from unconstrained deletion of arbitrary amount of source material between two sentences which end up adjacent in the summary; this has unpredictable effects on the amount of potentially essential information which may be lost in that deletion. Examples like 'dangling' anaphors (with lost antecedents) have been cited often enough, and strategies like including the immediately preceding sentence in the summary have some effect. While intuitively plausible, these are still 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.

We seek to leverage a mechanism for assessing the *degree of cohesion* between individual sentences in the source document, as well as having a notion of how these map onto the underlying themes in the document. Informally, cohesion—and *lexical cohesion* in particular—is manifest in the ways in which the words, or word patterns, of a sentence connect that sentence to certain of its predecessors and successors. The intuition is that identifying, and preserving, some of these connections in the summary would improve its coherence.

1.1 Lexical cohesion and summarization

Documents are coherent because of the continuity of their discourse. A number of rhetorical devices help achieve cohesion between related document fragments. Analysing such devices—or at the very least being sensitive to their manifestation and interplay—can bring a moderately refined degree of discourse awareness into the summarization process. In the absence of deep text understanding, this boils down to making extensive use of a formalized notion of lexical cohesion.

Linguists have studied extensively how various cohesive devices operate, and interact, in order to account for certain properties of the overall organization of a text discourse. For (Halliday and Hasan, 1976), the organization of text derives from a variety of relationships (*cohesive ties*) among discourse entities. More recently, (Winter, 1979) has focused on the devices that enforce lexical relationships and connect a discourse fragment with other discourse fragments. The underlying theme here is that cohesion can be best explained in terms of how repetition is manifested across pairs of sentences. Repetition carries informational value— it provides a reference point for interpreting what has changed, and thus, what is at the focus of attention of the discourse—and thus clearly goes well 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; this makes it possible to get a handle on the overall organization of text, in general, and on the identification of *topic introduction* and *topic closure*, in particular.

A variety of linguistic devices act as vehicles for repetition: viewed at the level of interplay 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. Analysing these would enable the identification of strong cohesive ties pulling together a chain of sentences which focus on (aspects of) the same discourse entity or event; this would require carrying out, for instance, in-depth co-reference and ellipsis resolution, as well as lexical relation determination.

At the other end of the spectrum, just a lexical chaining procedure (like the one described in (Morris and Hirst, 1991)) could be used to determine the degree of cohesion between adjacent pairs of sentences. Indeed, this has been the basis for an operational definition of linear discourse segmentation, where segments in a document are defined to be contiguous blocks of text, roughly 'about the same thing', with segment boundaries indicative of topic shifts.

The research reported here is just one aspect of a larger study into the recognition and use of cohesive devices for content characterisation tasks. It presupposes fine-grained methods for the identification of cohesive ties between (sentence) units in a text; describing the computational basis for developing such methods is outside of the scope of this paper (however, see (Kennedy and Boguraev, 1996), (Fellbaum, 1999), (Keller, 1994)), as is the complete framework for lexical cohesion analysis we have developed. Instead, in focusing on the effects of lexical cohesion on summarization, we limit ourselves here on the phenomenon of simple lexical repetition; it turns out that even this can be beneficially applied to enhancing summarization quality.

Recent work (Barzilay and Elhadad, 1999) makes explicit this intuition. “Lexical chains” are constructed by grouping together items related by repetition and certain lexical relations derived via the WORDNET lexical database (Fellbaum, 1999). A sequence of items in a chain highlights a discussion focused on topic related to (an) item(s) in the chain; a metric for scoring chains picks topically prominent ones; these are then taken as the basis of sentence extraction heuristics. A positive result of that work is that in an intrinsic evaluation against human-constructed summaries, the system outperformed at least one commercial summarizer. This highlights the potential of a purely lexical chains-based approach; still, Barzilay and Elhadad remain frustrated by the high degree of polysemy in WORDNET (not to mention its limited coverage with respect to more specialized domains); fortunately, this does not concern us here.

1.2 Discourse segmentation and summarization

Unlike Barzilay and Elhadad, we start with a sentence-based summarizer, and are specifically seeking to improve upon what is already (by some measure; see Section 4.1 below) a good performance, judged in a discipline-wide evaluation initiative (Mani et al., 1999). This places certain constraints on how lexical cohesion analysis results, and in particular the identification of topically coherent segments, can be incorporated in the existing strategies and mechanisms for sentence selection, already deployed by the summarizer. Making certain that a summary incorporates sentences from each segment intuitively seeks to ensure uniform representation of all sub-stories in a document; the notion here is to avoid having inordinately large gaps between adjacent summary sentences, which would tend to lose essential information. Moreover, a mechanism which would pick the sentence(s) in a segment most representative its main topic, would also carry over into the summary ‘traces’ of *all the main topics* in the original document.

This is more than just an intuition. In the process of developing, and training, our base summarizer (see Section 2.2 below), an analysis was carried out to determine the causes of a certain class of failure. It turns out that 30.7% of the failures could be prevented by a heuristic sensitive to the logical structure of documents, which would enforce that each (topical) section gets 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, leading / trailing noise (see 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 exemplifies how a document-wide analysis of a single lexical cohesion factor (simple repetition) can improve upon an existing sentence selection strategy—even if such a strategy has been devised *without prior knowledge* of additional enhancements to come. The specific approaches to being sensitive to foci of attention within a segment, and topic shifts between segments, may vary; as we discuss this below (see Section 3.1), these will depend on other environment settings for the summarizer. Still, in the right operational environment even very simple heuristics—take the first sentence from each segment, for instance—have remarkably noticeable impact.

We thus argue that a lexical repetition-based model of linear segmentation offers effective schemes for deriving sentence-based summaries with certain discourse properties, enhancing their quality.

What follows is organized in three main sections. We outline some linguistic functions of the summarizer, and give details of the summarization and segmentation components. We focus specifically on how higher level content analysis uses lower level shallow linguistic processing, both to obtain a richer model of the document domain, and to leverage cohesion analysis for sub-story identification. Next we discuss some strategies for optimal use of discourse segments and topic shifts for summarization. We sketch our evaluation testbed environment, and present experimental results comparing the performance of summarization alone to segmentation-enhanced summarization. We conclude 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.

2 Technology base

As an integral component of an infrastructure for document analysis with a number of interconnected and mutually enabling linguistic filters, the summarization system discussed here makes use of ‘shallow’ linguistic functions. The infrastructure is designed from the ground up to perform a variety of linguistic feature extraction functions, ranging from single pass tokenisation, lexical lookup and morphological analysis, to complex aggregation of representative (salient) phrasal units across multi-document collections. Given such a document processing environment, the design of our summarizer is based on sentence selection mechanisms utlilizing salience ranking of phrasal units in individual documents, when viewed against a background of the distribution of phrasal vocabulary across a large multi-document collection.

2.1 Linguistic filters

In essence, we have a robust text analysis system for identification of proper names and technical terms, since these are most likely to carry the bulk of the semantic load in a document. However, in addition to simple identification of certain phrasal types, capabilities also exist for identifying their variants (contractions, abbreviations, colloquial uses, etc.) in individual documents in a multi-document collection. A collection vocabulary of canonical forms and variants, with statistical information about their distribution behaviour, are used in the summarizer’s salience calculation. Salience, in turn, is a major component of the sentence-level score that selects the sentences for extraction (see 2.2 below).

As a frequency-based system, our summarizer is ideally positioned to exploit linguistic analysis, filtering, and normalization functions. Morphological processing allows us to link multiple variants of the same word, by normalizing to lemma forms. Proper name identification is enhanced with context disambiguation, named entity typing, and variant normalisation; as a result the system's frequency analysis is more precise, and less sensitive to noise; ultimately, this leads to more robust salience calculation. Normalisation 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; 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 collection is analyzed individually. All 'content' (non-stop) words, as well as all phrasal units identified by the 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 Section 5), these retain complete contextual information about the variants they have been encountered in, 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, 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. A pronominal anaphora resolution function further contributes to the quality of the collection statistics.

In addition to the domain vocabulary, the summarizer also has access to *document structure* information. A hierarchical representation of the document separates content and layout metadata, and makes the latter explicit in a document structure tree. Encoded are data including: appearance and layout tags; document title; abstract, and other front matter; (sub-)section, etc. headings; paragraphs, themselves composed of sentences; 'floating' objects like tables, figures, captions; side-bars and other text extraneous to the main document narrative; etc. Document structure is constructed by 'shadowing' markup parsing, as markup tags are used to construct the document structure tree; for documents without markup, structure determination is carried out on the basis of page layout cues. The document structure records additional discourse-level annotations, such as cue phrases marking rhetorical relations, quoted speech, and so forth. All of these elements both contribute directly to the summarizer's set of heuristics, as well as inform the discourse segmentation process.

2.2 Salience-driven summarization

With its set of linguistic filters, our frequency-based summarizer can exploit linguistic dimensions beyond single word analysis; this is not unlike the approach of (Aone et al., 1997). Due to the sophistication and integration of the filters (see Section 2.1), we are able to exploit a richer source of domain knowledge than most other frequency-based systems.

Frequency alone is poor indicator of salience, even when ignoring stop words. Unlike early frequency-based techniques for sentence selection, we utilize the more indicative *inverse document frequency* measure, adapted from information retrieval, in which the relative frequency of an item in a document is compared with its relative frequency in a background collection. The trade-off, however, for more precise term salience is the summarizer's dependence on background collection statistics; we return to this issue below.

Sentence selection is driven by the notion of salience; the summary is constructed by extracting the most salient sentences in the full document. The *salience score of a sentence* is derived partly from the salience of vocabulary items in the document 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 calculation of inverse document frequency for a vocabulary item t compares its relative frequency in the document with its relative frequency 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).

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

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.; see Section 2.1). 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 paragraph, and its 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 address some of the coherence-related problems discussed earlier (see 1). For example, under certain conditions, a sentence might be selected for inclusion in the summary, even if it has low, or even zero, score: sentences immediately preceding higher scoring ones in a paragraph may get promoted by virtue of an 'agglomeration rule'. Agglomeration is an inexpensive way of preventing dangling anaphors without having to resolve them. Another problem for sentence-based summarizers, that of thematic under-representation (or, loosely speaking, coverage; see 1), is addressed by an 'empty section' rule, which is of particular interest for this paper. Longer documents with multiple sections, or news digests containing several stories, may be unevenly represented in a sentence-extracted summary. The 'empty section' rule aims to ensure that each section is represented in the summary by forcing inclusion of its highest scoring sentences, or, if all sentence scores are zero, its first sentence.

As a general purpose summarizer, ours makes extensive use of small scale linguistic information (term phrasal patterns) and large scale statistical information (term distribution patterns). 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 departure point for this work, as already discussed (in Section 1): can the summarizer's performance be improved, if we take into account lexical cohesion in the source?

We address this question 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 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* (see Section 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 take into account multi-word sequences (proper names, technical terms, and so forth, as discussed in Section 2.1 above), as well as a 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—are exploited for their 'pre-disposition' to act as likely segment boundaries; and so forth (see Section 2.1). 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 our document processing environment, we are able to use, transparently, the results of processes such as *lexical and morphological lookup, document structure identification*, and *cue phrase detection*. Likewise, segmentation results are 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 Hearst's formula for computing lexical similarity between adjacent blocks of text b_1 and b_2 (t denotes a discourse element term identified as such by 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):

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

Unlike most applications of segmentation to date, which are concerned with the identification of segment boundaries, we are primarily interested in leveraging the content of the segments, to the extent that it is indicative of the focus of attention, and (indirectly, at least) points at the topical shifts to be utilized for summary generation. We use the segmentation results (together with the name and term identification and salience calculation delivered by other functions) in order to ensure that all the base data for inferring the topic stamps, and topic shifts, is available to the user.

3 Segmentation-assisted summaries

What is the relationship between segmentation and summarization: is segmentation a strictly "under the covers" function for the summarizer, or might segmentation results be of any interest, and use, to the end user? We focus on some strategies for incorporating segmentation results in the summary generation process. However, unlike (Kan et al., 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 in-depth perception of the summary, as it relates to the full document (Boguraev and Neff, 2000).

3.1 Strategies for utilizing segments

Common intuitions suggest a number of strategies for leveraging the results of linear discourse segmentation for enhancing summarization. As topic shift points in the text are 'published' into the document structure (see Section 2.3), by defining a segment as an additional type of document span (akin to sentence, paragraph, section, and so forth), the summarizer transparently, and immediately, becomes aware of the segmentation results. We also make arrangements for a mechanism whereby certain strategies for incorporating segmentation results into the summarization process were easy to cast in summarizer terms.

Thus, for instance, a heuristic requiring 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). The selection of a segment-initial sentence for the summary can be enforced 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.

3.2 Other benefits of segmentation

Such strategies are discussed in more detail later, 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 sentence-based summaries with certain discourse and rhetorical properties, it turns out that under certain conditions the summarizer can operate very effectively without a need for background corpus statistics. This is a better solution than the highly genre-sensitive 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. Note that the derivation of a background collection and statistics for it might be impractical for a variety of reasons: lack of access to a sufficiently large and representative data sample; no time for processing; sparse storage resources; and so forth. Clearly, being able to operate without such statistics is an operational bonus for the summarizer.

Another use for segmentation is for optimising the use of source input, as well as possibly maximising its re-use. Occasionally, the document contains 'noise'—possibly in the form of *anecdotal leads*, *closing remarks* tangential to the main points of the story, *side-bars*, and so forth—which are inappropriate sources 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 news reporting genres a whole document fragment (typically towards the beginning or the end of the document) functions as a summary of the story: we would like to be able to use this fragment; clearly identifying it as a segment would help.

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—see e.g. (Mitra et al., 1997)—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

For evaluating the effect of various strategies upon summarizer output quality, we used 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 whether task-based evaluation is appropriate for testing strictly the effect of one technology on another (see Section 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.

The remainder of this section describes our evaluation environment, and then looks at 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 correct, 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. *Subjective* measures, even though still less satisfying, can also be devised: for instance, summary acceptability has been proposed as one such measure. 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 et al., 1995); or recall, precision, and time required in document categorization (i.e. assessing whether a document has been correctly judged to be relevant or not, on the basis of its summary alone) (Mani et al., 1999).

We built an environment for baseline summarizer evaluation, as part of its development/training cycle. This was also used in analyzing the impact of discourse segmentation on the summarizer's performance. A background collection vocabulary statistics was derived from analyzing 2334 *New York Times* news stories. Sentences in digests for 808 stories and feature articles were automatically matched with their corresponding sentences in the full-length documents. 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.

Since 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 linguistic 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 Section 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); *4 sentences*; *10% of the size* of the full length document; and *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 (Section 2.2), with enhanced procedures incorporating different strategies using the notions of discourse segments and topic shifts.

These elaborate the intuitions underlying our approach to leveraging lexical cohesion effects (see Section 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% (Section 4.1).

We experimented with two broad strategies for incorporating topical information into the summary. One approach aimed to bring ‘topic openers’ into the summary, by adding segment-initial sentences to those already selected via salience calculation. The other was to exert finer control over the number of sentences selected via salience, and ‘pad’ the summary to its requested size with sentences selected from segments by invoking the ‘empty segment’ (aka ‘empty section’, see 2.2) rule. Special provisions accounted for the fact that segmentation would naturally always select the document-initial sentence.

It turns out that the differences between a range of realisations of the above two strategies are not statistically significant over our test corpus; we thus use the label “SUM+SEG” to denote a ‘composite’ strategy and to represent the whole family of variations. In contrast, “SUM” refers to the base summarization component, and “SEG” represents summarization by segmentation alone. Table 1 below shows the recall rates for the three major summarization regimes defined by different summary granularities. 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 encapsulates a ‘composite’ result (denoted by the label “*All documents*”).

	4 sents	10%	20%
All documents			
SEG	54.74	54.74	56.09
SUM	46.85	49.71	66.47
SUM+SEG	56.52	56.30	58.37
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
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

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. “*All documents with* $>$ *1 digest sentence*” represents documents whose digests are longer than a single sentence; “*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.

Overall, leveraging some of the segmentation analysis is positively beneficial to summarization; the effects are particularly strong where short summaries are required. In addition, summarization driven by segmentation data alone shows recall rates comparable to, and in certain situations even higher than, the baseline: 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 Conclusion

Starting from a class of problems inherent to summarization by sentence extraction, 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 (calculated 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. 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

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

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

Steven Bird 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.

Branimir Boguraev and Mary Neff. 2000. Lexical cohesion, discourse segmentation and document summarization. In *Proceedings of RIAO-2000, Content-Based Multimedia Information Access*, Paris, France.

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

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

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

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

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

Min-Yen Kan, 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.

Andrew Keller. 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.

Christopher Kennedy 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.

Inderjeet Mani, 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.

Seije Miike, 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.

Mandar Mitra, 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.

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

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

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