IBM Research Report

Summarization Miniaturization: Delivery of News to Hand-Helds

Branimir K. Boguraev, Rachel K. E. Bellamy, Calvin B. Swart

IBM Research Division

Thomas J. Watson Research Center

P.O. Box 704

Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Haifa - India - T. J. Watson - Tokyo - Zurich

Summarisation Miniaturisation: Delivery of News to Hand-Helds

Branimir Boguraev and **Rachel Bellamy** and **Calvin Swart**

IBM T.J Watson Research Center, 30 Saw Mill River Road, Hawthorne, NY 10532

E-mail: {`bran,rachel,cals`}`@us.ibm.com`

Abstract

The networkability and ubiquity of hand-held devices makes them ideally positioned to mediate access to on-line news. Existing frameworks for news delivery to hand-helds typically involve transcoding methods by remote proxies layered over generic summarisation techniques. We argue that such environments, characterised by both analysis and presentation being done on a server, fail to account for essential features of the news delivery client. This work addresses two questions in this context: the specialised 'transcoding' strategy for a well-defined sub-type of content, namely that of primarily text-based news documents, and the emergence of a 'summary-for-a-hand-held' genre, which exploits advanced linguistic analysis to meet the particular requirements of information seeking by news skimming on hand-helds. Directly related is the issue of how novel methods for deriving context- and profile-sensitive document abstractions interact with novel metaphors for mediating these abstractions according to the particulars of their use.

1 Information seeking and hand-held devices

In this paper we describe a framework for viewing on-line news on wireless hand-held devices. Hand-held devices present a significant departure from established modes and metaphors for information access and management. Originally conceived as personal information managers (handling calendar appointments, diary entries, to-do lists, ephemeral notes, and so forth), they are rapidly gaining wider deployment. Moreover, they are being targeted by a wider class of applications, which view wireless hand-helds as generic clients for custom information delivery.

"Custom", so far, typically tends to mean either designing specialised applications (such as, for instance, the interface for a home shopping application (Bellamy et al., 2000)), or adapting content for hand-held delivery by means of deploying transcoding proxies between a generic content pro-

viding server and the hand-held client (Smith et al., 1998). In any case, such applications present significant challenges for the design of electronic formats and interfaces to them, given that hand-held devices as a medium have limited display and interaction capabilities. Their screens are small; (to date, at least) not all of them support colour; those that do face some limitations, compared to desktop colour functionality; and input to them is typically by pen (or, on occasions, by voice).

Still, hand-helds have several advantages over desktop machines and large screen displays, not the least of which is that they can be carried around at all times. This makes them highly personal devices, reflecting their users' profiles; a significant consequence of this is that they get to be deployed in very diverse contexts and environments, compared to, say, desktop machines. Consequently, the notion of delivery of electronic documents content, for information seeking purposes, via hand-helds needs rethinking.

Two questions in this context are addressed in this work: the specialised 'transcoding' strategy for a well-defined sub-type of content, namely that of primarily text-based news documents, and the emergence of a 'summary-for-a-hand-held' genre, which exploits advanced linguistic analysis to meet the particular requirements of news skimming for information seeking on hand-helds.

1.1 Summarisation and transcoding

Scalable, multimedia, delivery for pervasive devices is a rapidly growing area of research: see, for instance, (Mohan et al., 1999), and citations therein. Primary focus of that work is reduction and/or compression of content so that e.g. a Web page can be displayed on a palm-top.[1]

Typically, for pages fetched from a server, a transcoding proxy compresses and reformats their content to match the capabilities of the client. Compression and reformating are defined as different

---

[1] Some hand-helds—for instance those running Windows CE—offer a broader range of functions than the generic Palm Pilot, often on superior hardware. However, given the wider spread of PalmOS to date, this paper focuses on palm-top clients.

operations over a range of different content type and client characteristics. In principle, a proxy approach advantageously allows content authoring to be unaware of the client specifics. In practice, the relatively straightforward nature of the content type we are interested in here—news articles which typically consist of a linear text stream with perhaps an image or so embedded within it—makes such an advantage somewhat of an irrelevance.

For news delivery on a hand-held, techniques for summarisation/content abstraction of electronic documents become critical enabling technologies. Current strategies for transcoding of news pages achieve text condensation simply by deploying a summarisation engine. The prevailing methods for summarisation, however, primarily sentence-extraction based, suffer from a number of drawbacks which get amplified by the limitations of hand-helds. The diverse social and personal contexts in which users might want to use news highlights add further constraints on current summarisation techniques.

In this paper, therefore, we focus on a content analysis technology which is specifically designed to scale down to fit the limited display capabilities of the platform. The summarisation function built on top of this technology exhibits properties tailored to meet the requirements of a new 'summary-for-a-hand-held' genre.

It is *less rigidly sentence-based*, in that its sentence selection methods have a more acute sense of context. It has *a flexible notion of granularity (size) of an information-bearing nugget*, by focusing on sub-sentence level fragments (such as phrasal constituents or clausal units) with topical significance. It also seeks *to adjust the size of its summaries dynamically*, to optimally fit a hand-held screen, by means of a capability to compress sentences without losing readability. In addition, it makes allowances for the highly personalised and individualised nature of hand-helds, by enabling readers to view multiple threads in a story, ultimately allowing for multiple perspectives into the story.

We also take the view that instead of transcoding a generic summary, by a proxy remote from the client, more informative delivery of content highlights can be realised by explicitly taking into account the client specifics, and feeding these back as parameters into the linguistic analysis environment. Additionally, we take into account that some of the summary (re-)formating for rendering on the palm-top would best be done on the client.

2   News stories and small screens

In essence, an environment like the one we are developing requires novel methods for deriving profile- and context-sensitive document abstrac-

tions interacting with novel metaphors for mediating these abstractions on the basis of who, where, and when is using them.

Most summarisers in broad deployment today generate a summary as a concatenation of representative fragments, typically sentences, which have been determined to have high degree of representativeness with respect to the main theme(s) of the source document. As a strategy of sentence selection which simply picks sentences above a certain threshold level, this exhibits a particular fault: it is incapable of being sensitive to topic shifts; nor can it follow specific topic threads (Boguraev and Neff, 2000).

There is no such thing as best, or canonical, summary (Spärck Jones, 1993). Different users may seek, and react to, different aspects of the same document content. Thus, a single document is potentially of interest to many people for many reasons; also, in different contexts, alternative perspectives on the information in the document might be required. A generic summary of a document, while allowing for quickly absorbing content highlights, may well miss aspects of the information of particular interest or relevance to any single reader

Generic summarisers are not at all sensitive to such issues; arguably, they do not have to be, as they tend to be deployed at the server end of a Web portal, and have very little notion of the individual users who might be accessing their summaries.

In contrast, presentation of document highlights for hand-held devices must take into account the individuality of the particular user. A 'brute force' approach to meeting the requirements of a broad range of users—such as overgeneration in order to cater for multiple points of view—is inappropriate here, as the physical limitations of the device constrain heavily the size of a summary. It is essential that content abstraction techniques be developed which are capable of *multiple perspective summaries*, each coherently following a given aspect, or a set of related topics, through the same source document.

Furthermore, when viewed from the perspective of a hand-held client, generic summarisers exhibit other characteristics which make the notion of client-detached transcoding (Mohan et al., 1999), at best, inappropriate for content highlights delivery. The most egregious of these are to do with the notion of a sentence being an indivisible unit of content, about which no knowledge (structural or semantic) exists, and which cannot be readily judged in relation to other sentences. These are natural consequences of the fundamentals of sentence selection strategies (Mani and Maybury, 1999); and in environments where summaries are delivered on desktop machines, enjoying large displays, substantial processing power, and advanced graphical

capabilities, the jaggedness of a summary 'stitching together' source sentences can be softened by apposite visualisation techniques (Boguraev et al., 1998).

However, let us consider, for instance, the popular Palm V personal organiser. With its 160x160 pixel-sized screen, it requires well over 20 screenfuls to display a text document which is of the order of two print pages. This is clearly too much for smooth (and quick) reading of a news story, and further demonstrates the need for a summarisation function, specially tailored for the Palm.

While the relationship between the specifics of the user task and interface characteristics is complex (Pirolli and Card, 1998), studies show that when browsing for information seeking, a close coupling is observed between the access cost of information and the propensity for it being used (e.g. read or cited) (Soper, 1976). Thus, in designing a tool for news delivery, it makes sense to reduce the number of interface actions that must be made in order for the reader to get the gist of the story and know whether it is of interest to them.

For small devices such as the Palm specifically, reducing the number of screenfulls to scroll through for the reader to get the gist of a document is one way to immediately achieve a reduction in interface actions. Generic sentence-based summarisers tend to generate reductions of the order of 25–30% of the original. This means that a summary of the above document would require 7 or more screenfuls to display; indeed, a summary generated by our production system (Boguraev and Neff, 2000) running with default settings spans almost 8 Palm screenfuls (Figure 1); this is a considerable contrast in form-factor terms, when compared to the single page/screen sufficient to display the same summary on a desktop machine.
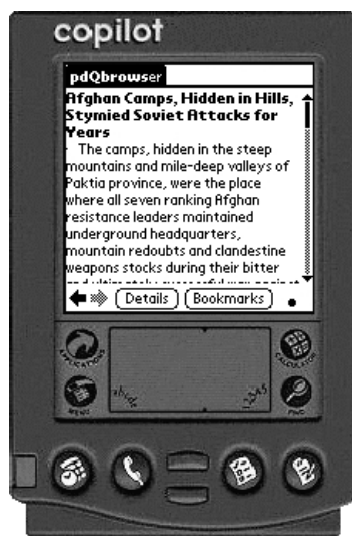

Figure 1: A generic summary of a news story

Clearly, the reduction strategies at the desktop, by large involving varying the threshold of sentence selection and dropping additional sentences from the summary, are inappropriate for hand-helds. In addition to the need for active scrolling, several other characteristics of the display illustrated in the figure are immediately obvious. For instance, the title takes up too much real estate; and yet, if we want it typeset for some visual impact, we have no choice but to use the larger font. Furthermore, it is not uncommon for sentences in news stories to be complex and long—note the opening sentence, which needs over a screenful by itself—which makes fine adjustment in reduction factor impossible.

Typically, such adjustments are defined in terms of specifying target size of the summary as a percentage of the original document; the operation adds, or removes, whole sentences to the summary. Generic summarisers are relatively robust with respect to omitting (or adding) one or more additional sentences: since selection is based purely on a ranking scheme where individual sentences are assigned a score in isolation, there are no predictable effects from moving the cut-off point higher (or lower) in that ranking. However, as we argue below, more appropriate to the highly personal nature of hand-held devices is a procedure for identifying and following particular story threads, by generating summaries based on sentence chains with high degree of inter-sentential cohesion. Since dropping one or more sentences from such a chain would degrade its utility as a summary representation, it is clear that an alternative mechanism for summary size reduction is required.

3 Multi-perspective, scalable summarisation

We have developed strategies for deriving summaries tailored to some of the requirements discussed in the previous section. In particular, our methods are *sensitive to topic threads*, and thus support the capability of viewing the same document from different perspectives. They are also *scalable*, in the sense that a summary can be 'shrunk' in size by dynamically adjusting the length of summary sentences, rather than fully eliminating sentences. The strategies complement each other, and are fundamentally enabled by the same set of underlying linguistic functions.

3.1 Story lines and summaries

Our strategy for deriving multiple summaries from a single document, representative of different story lines and reflecting different topical perspectives, is based on analysing a number of phenomena contributing to the *cohesion* of a text document. In particular, we seek to leverage a mechanism for as-

sessing 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 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.

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 summarisation process. By analysing several cohesive factors, we determine connectedness between text fragments (typically sentences and/or paragraphs); this in turn is used to derive cohesive (and coherent) substory threads.

Linguists have studied extensively how various cohesive devices operate and interact. For (Halliday and Hasan, 1976), the organisation of text derives from a variety of relationships (*cohesive ties*) among discourse entities. More recently, Winter (1979) has focused on the devices that connect a discourse fragment with other discourse fragments. The lexical inventory of a text is tightly organised in terms of repetition and collocation (Phillips, 1985); this makes it possible to get a handle on the overall organisation of text, in general, and on the identification of *topic introduction* and *topic closure*, in particular (Hoey, 1991).

3.2 Cohesion chains

In essence, text cohesion is best explained in terms of how *repetition* is manifested across pairs of sentences. Repetition carries informational value—it is a reference point for interpreting what has changed, and therefore, 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.

Numerous 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.

Summarisation research has actively explored such devices for deriving intermediate representations for summaries. Notably, Barzilay and Elhadad (1999) pioneered a purely lexical chains-based approach to summarisation, exploiting only lexical repetition and lexical relations. Issues of high degree of polysemy in the lexical resource used (WORDNET, (Fellbaum, 1999)) and computational complexity of their approach have more re-

cently been addressed with highly encouraging results (Silber and McCoy, 2000). Lexical cohesion has also been used in summarisation tasks such as cross-document summarisation (Mani and Bloedorn, 1999), query-biased summarisation (Okumura et al., 1999), and salience-based summarisation (Boguraev and Neff, 2000). Another cohesion factor, co-reference, has also been exploited in the summarisation context; for a representative sample see (Baldwin and Morton, 1998), (Azzam et al., 1999), and (Boguraev and Kennedy, 1999).

In the spirit of such work, our summarisation procedure incorporates, and is equally informed by, cohesion analysis mechanisms which automatically identify and use lexical repetition, lexical relations, and intra- and inter-document co-reference. Strong cohesive ties highlight and connect a chain of sentences which focus on (aspects of) the same discourse entity or event. Typically, more than one such chain can be identified in a document, corresponding to how subsets of salient discourse objects collocate across the entire document.

While there are still some cohesion factors which are hard to analyse computationally—*ellipsis, paraphrasing, conjunction*—the inventory of analysable ones outlined above is sufficient to enable a scoring procedure for identifying lexical chains through a document. The document processing environment we are using integrates a broad range of interconnected, and mutually enabling, linguistic filters; these are designed from the ground up to perform a variety of linguistic feature extraction functions including: lexical look-up and normalisation, morphological analysis, named entity identification (with abbreviation and coreference analysis) and aggregation, phrase- and clause-level analysis, anaphora resolution, and salience calculation (Boguraev, 2000). These capabilities underpin an infrastructure for cohesion analysis of the kind detailed above.

3.3 Cohesion chain analysis

For each pair of sentences in the document, a base-level 'connectedness' parameter is initialised. The document stream is broken down into word tokens, and phrase sequences over those (see below for details concerning phrase recognition). The tokens are analysed for lexical repetition and morphological relatedness; if two tokens in different sentences are identical (lemma identity) or morphologically related, the 'connectedness' value between the sentences is incremented.

Independently, a shallow parser (described in some detail in Section 3.5) recognises phrase and clause units of certain types, over which a cascade of cohesion analysis procedures operate. Named entity identification looks for proper names and
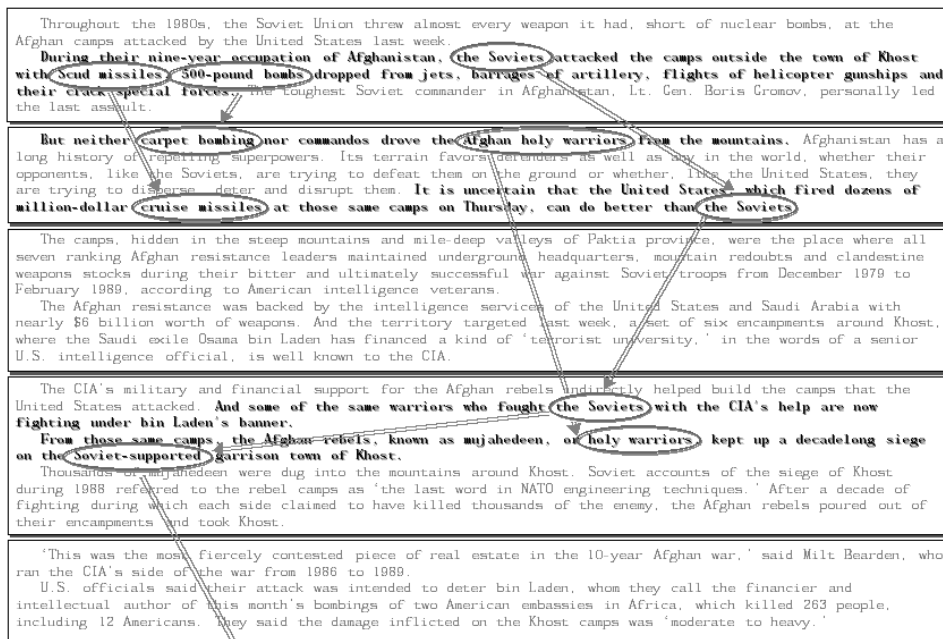
Throughout the 1980s, the Soviet Union threw almost every weapon it had, short of nuclear bombs, at the Afghan camps attacked by the United States last week.
   During their nine-year occupation of Afghanistan, the Soviets attacked the camps outside the town of Khost with Scud missiles, 500-pound bombs 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: Summary derived from cohesion chain *"Soviets"*, *"holy warriors"*, *"bombs"* and *"missiles"*

their variants (contractions, abbreviations, and so forth), and is followed by named entity coreference module. Noun phrases are categorised (e.g. simple, complex, embedded, descriptive, named entities, appositive, lists, and so forth) and definite noun phrase anaphora establishes links between the original mention of an object, and subsequent references to it. Similarly, although running as a separate process, coreference chains are extended to account for pronominal anaphoric references. As an example, a coreference chain might incorporate the phrase initially introducing an object in the discourse (*"Osama bin Laden"*), a contracted form encountered later in the document (*"bin Laden"*), and one or more anaphoric references to it (*"he"*, *"his"*).

The categorisation of different phrasal types, as identified by the shallow parser, makes it possible to localise tests for relatedness to syntactic constituents of the same type—so that nouns are compared with nouns, verbs with verbs, and so on.

As a side effect of coreference resolution, a *salience* value is calculated for each discourse referent; this is intended to indicate the discourse prominence of the referent. Actual values are not of importance; rather, salience values are used to rank discourse referents by some notion of (global) relevance, making it possible to compare different discourse objects with respect to their prominence in the document (Boguraev and Kennedy, 1999).

Coreference chains additionally boost the 'connectedness' parameter between sentences which contain coreference chain members. Similarly, other lexical relatedness tests (such as, for instance, for synonymy) further constribute to the overall connectedness parameter for the sentence pair containing the synonymous items. The notion is to define, and calculate, a numerical value for each pair of sentences in the document, which acts as 'strength-of-cohesion' indicator between the two.

3.4 Story threads

Once the individual cohesion factors analysis has been completed, sentence pairs with connectedness value below a certain threshold are discarded from consideration. Viewing what remains as a directed graph, where a sentence may be connected with one or more preceding and/or following sentences, we are now in position to identify threads through the document which are defined by prominent repetiton of identifiable subsets of topically relevant discourse referents. A document may well hold more than one such thread, as different aspects of discourse referents get elaborated in one or more, not necessarily contiguous, sentences.

Schematically, this kind of analysis is illustrated in Figure 2. (For simplicity, not all cohesive connections between sentences are displayed, and only the opening few paragraphs are visible.) The diagram shows some of the cohesion chains connecting discourse referents: *"Scud missiles"* and *"cruise missiles"*, *"500-pound bombs"* and *"carpet bombing"*, *"the Soviets"* (repeatedly) and *"Soviet supported"*, *"Afghan holy warriors"* and *"holy warriors"*. Some chains will extend across the entire span of the document; again, they are truncated here for simplicity.
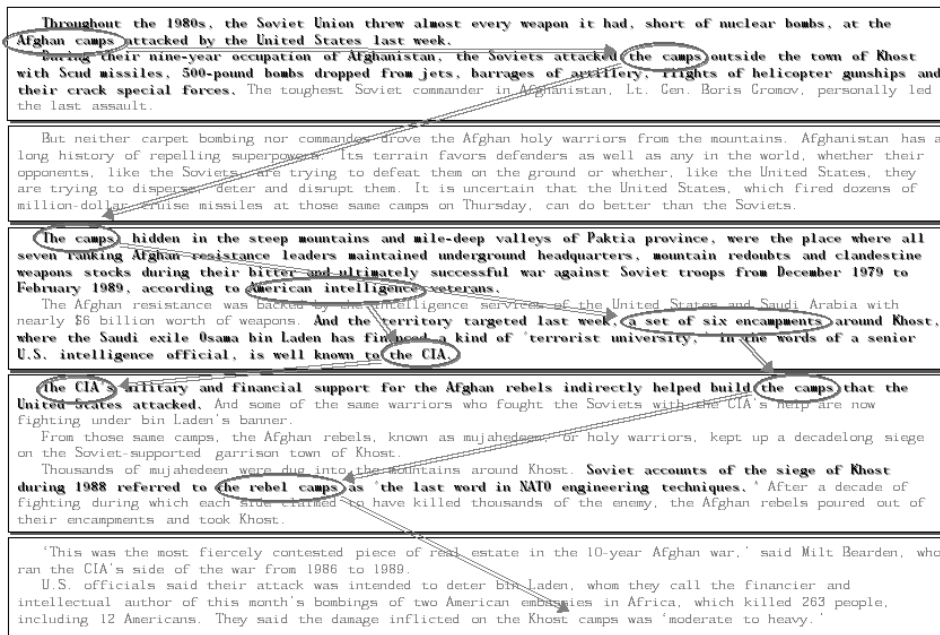
Figure 3: Summary derived from cohesion chain *"camps"*, *"encampments"*, *"American intelligence"* and *"CIA"*

Accumulating the effects of establishing these connections, cohesive ties within certain sentence pairs boost their 'connectedness' score above a threshold. In essence, this means that any pair of sentences connected by virtue of more than one cohesion chain running through, share one (or more) discourse referents. The two sentences thus are giving complementary detail concerning these discourse objects, and joining them together—even if this might omit some intervening material—would result in a moderately coherent passage, because by definition there are going to be shared referents in these sentences.

Following a chain of sentence connections established in such a way, we derive a thread, or a story line, through the document which acts as a summary from a particular perspective. In the above example, the summary (composed of sentences highlighted in bold: one in the first segment, two in the second segment, one in the fourth, and so forth) highlights the interactions between (Afghan) holy warriors and their Soviet adversaries, with additional detail concerning missile and bomb engagements.

Other cohesion chains also run through the same document, and some of them are sufficiently strong to highlight another set of closely related topical items. Figure 3 illustrates a different configuration of sentences, defining a different thread through the document, licensed by connections established by repetitions of discourse objects focusing this time on *"camps"* (*"Afghan camps"*, *"rebel camps"*, *"a set of six encampments"*) and *"American intelligence"* and

"the CIA". This story line gives rise to a different summary. There is some overlap with the earlier example, as sentences may incorporate discourse referents from the two different topical sets. This is consistent with our approach of focusing on cohesion chains, and deriving sentence chains which incorporate elaborations for all the topical members of a lexical chain.

The two figures illustrate, indirectly, that this method is also sensitive to topic shifts. A separate process of linear discourse segmentation (Boguraev and Neff, 2000) has identified the points in the document where topic shifts occur; these are schematically shown in the figures grouping paragraphs into larger cohesive segments. As Figure 2 shows, the story line 'bypasses' the third segment (which discusses Afghan resistance camps and their backing by US intelligence). Similarly, the thread in Figure 3 skips over the second segment (which discusses the history of Afghanistan's engagement with different superpowers).

3.5 Summary compression: sentence reduction

The distinctive story lines in a document derived by the processes described above are defined in terms of sequences of sentences; by definition, there is higher than average degree of cohesion between these sentences. Such summaries will be used in an environment where their aggressive compression is the norm. Still, as we discussed in Section 2, modulating summary size by selectively removing sentences from it is an inappropriate strategy, given our method for sentence selection which exploits

the contiguity of a cohesive chain.

Here we describe an alternative process of summary compression, which exploits the notion of sentence reduction. As we saw (Figure 1), sentences in news stories tend to be long and complex, and it is not uncommon for these to be too long to fit on a single screen. This complexity is a feature of the news genre: it arises from attempts, on the part of the author, to address numerous aspects of one or more story lines; thus while clearly contributing, directly, to sentence length, complexity also allows us to consider dropping (deleting) certain fragments of a sentence, within the boundaries of retaining grammaticality, and assuming that we do not lose relevant (from the perspective of a given story line) topic descriptions and elaborations.

Our sentence reduction process is based on two intuitions. In principle, the configurational properties of a sentential syntactic structure offer a relative ranking of prominence of different phrasal and clausal constituent units, on the basis of which some deletion of material of 'secondary importance' could be performed. Additionally, as a result of the cohesion analysis process outlined in Section 3.3, we have salience ranking for all the discourse referents in the document; this would allow us to make additional informed decisions concerning material which might be considered peripheral from the perspective of a summary following a particular story line.

The crucial element here is the assumption of refined syntactic analysis. While current state-of-the-art syntactic parsers are capable, on the whole, of performing broad syntactic analysis of arbitrary text, for reasons primarily to do with coverage and efficiency, such parsers cannot be deployed in the operational context of our real-time document processing environment. We do, instead, make heavy use of finite-state phrase recognition and composition techniques, which substantially emulate the complexity and coverage of large grammar based approaches (Boguraev, 2000). (In this, we share the approach outlined by Grefenstette (1998); however, instead of applying 'blindly' just a set of filters defined purely in terms of syntactic depth, we use the salience calculated for the discourse referents, as described above).

We use a cascade of progressively more complex finite-state grammars. This is designed to agglomerate simpler phrases recognised earlier into the cascade into more complex phrasal units, taking into account clause boundaries, and determining (partial) grammatical function for key constituents. Without going into details, we illustrate in Figure 4 the kind of analysis derived by such a cascade.

The primary function of such analysis is to identify phrasal units corresponding to objects and

```
[PP Throughout [NP the 1980s NP] PP] ,
[SUB [NP the Soviet Union NP] SUB] [VG
threw almost VG] [OBJ [NP every weapon
NP] OBJ] [MC [SUB [NP it NP] SUB] [VG had
VG] MC] , short [PP of [NP nuclear bombs
NP] PP] , [PP at [NP the Afghan camps NP]
PP] [MC [VG attacked VG] [PP by [NP the
United States last week NP] PP] MC].
```

Figure 4: 'Shallow' syntactic analysis

events in the domain, i.e. noun and verb groups. This is essential for the cohesion analysis, but also makes it possible to construct, subsequently, a much richer representation of more complex objects and events. Thus following the bracketing of *"Afghan camps"* and *"the United States"* as noun phrases (at the end of the sentence above), additional processing brings in some peripheral material, constructs enclosing larger (in this case prepositional) phrases, identifies clause boundaries and marks the clause trailing behind *"Afghan camps"* as modifying the NP it precedes. There is another modifier clause in the sentence, similarly marked MC. Note also the grammatical function tags, SUBJ and OBJ, also identified by finite-state methods.

Whether this kind of analysis is faithful to the deep relational structure of the sentence is a question which need not concern us here; we certainly would not assume it complete or correct to the extent that it would support faithful semantic interpretation. But configurational information derived by a finite-state cascade as illustrated here is adequately informative to support the kinds of text processing operations we have in mind, namely the cohesion analysis described earlier, and the sentence reduction heuristics which could be described simply as following the intuition that secondary material is potentially droppable from the sentence.

The operational definition of 'secondary' material is somewhat fluid. Anything that is not a subject, top level verb, or object could be considered as peripheral. The same applies to adjunct information to the main clause components; thus, a clause modifying an NP in subject or object position would be considered peripheral, just as trailing prepositional phrases would be too. The analyser explicitly labels different phrase and clause types, and there is an ordering of phrasal types in terms of their prominence (Section 3.2). This is in the absence of any extraneous contextual information; additional considerations might weigh for or against dropping a constituent from a summary sentence, depending on whether it is a part of the cohesion chain which licenses this particular sentence for inclusion

in this particular summary. Additional heuristics are defined for certain constituent types and configurations: thus an NPLIST might be reduced to its first contituent; appositive NPs might be reduced to their (syntactic) heads; NPs within PPs within NPs might be deleted; and so forth.

Of particular importance to our application is that a cascade like the one described here can be implemented to run on the client, leaving to the server the task of deriving only the intermediate summary representation, and allowing the client to reconfigure the rendering of shorter or longer summaries, as requested by the user. It is clear that delegating such task to a transcoding proxy is impractical.

Unlike some related work (Jing, 2000), (Knight and Marcu, 2000), decisions concerning exactly how to compress a sentence are purely syntactically driven. This is motivated in part by the lack of suitably annotated corpus of sentence transformations in the context of delivering multi-perspective summaries (in other words, no suitable training data exists), and in part by the realisation that the same sentence might be reduced in different ways, depending on how it contributes to the story thread it participates in. For instance, the sentence in Figure 4 could drop trailing prepositional phrase(s); however, depending on whether it participates in a cohesion chain mediated by *"bombs"*, or by *"camps"* (see Figs. 2, and 3), we would need to retain either of *"short of nuclear bombs"* or *"at the Afghan camps attacked by ..."*.

Without contextual overrides, a condensed version derived from the syntactic configuration in Figure 4 would be: *"[...] the Soviet Union threw every weapon it had, [...] , at the Afghan camps [...]"*.

4 Summary delivery to hand-helds

A custom interface for delivery of multi-perspective summaries to a hand-held Palm Pilot is under development. However, it is instructive to view the results of the analysis via a 'bare bones' Web browser, which simply accesses Web pages with intermediate representations and displays them on the Palm screen.

A fundamental challenge for the design of a content page on a hand-held, exacerbated by a small screen, is the limited inventory of typefaces, and equally limited set of methods for in-line highlighting and spatial organisation. The interpretation of the HTTP 3.2 standard on the Palm Pilot (Qua, 1999), for instance, renders everything in one of two fonts, has no support for frames, and is very limited in interpretating tables markup.

In general, this means that content has to be displayed in normal type; highlighting by modulating the typeface is impossible; document titles will, almost inevitably, be too long. And yet, as we saw in

Figure 1, the larger, bold one is too intrusive to use on multi-line titles. Therefore, if we wish to retain the `<title>` markup, with its associated semantics, we are forced to apply radical compression on the title phrases (using the text reduction process described in the previous section). This is unfortunate, as titles are clearly highly indicative of content; on the other hand, it is indicative of the hard constraints of this operational environment. This is also another example of a requirement, at the client end, which is likely to be ignored—and certainly cannot be addressed explicitly—by a generic, remote, transcoding proxy (Section 1.1).

Another question is that of mediating the different story lines/summaries to the user. Even if now we have the capability to offer highlighting of certain aspect(s) of a document, it is still the case that the users have to be made aware of what the options are. In interface design terms, the problem is to express the topical highlights in the document, to show how each of these is a part of possibly more than one story line in the document, to show the 'gist' of a story line (for all the cohesive threads through the document), and to navigate from this to the corresponding summary. Again, none of the presentational and navigational metaphors for gisting and browsing conceptual spaces with complex interconnections (see, for instance, (Boguraev et al., 1998))—bi-directional scrolling, smooth zooming, mouse rollovers, synchronised displays—are read-
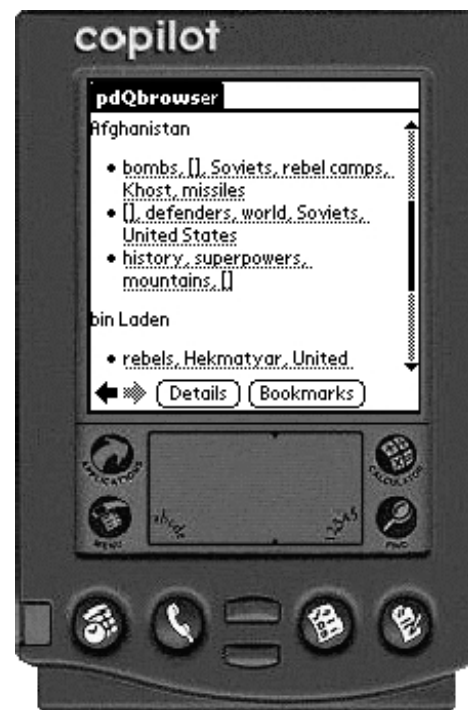

Figure 5: Topical highlights and summaries.

ily transferrable to the Palm.

Let us consider a screen shot of some of the information we are currently designing for. Figure 5 shows, after exploring the space of topic stamps for a document (in this particular example, this is a list containing the phrases *"Afghanistan"*, *"bin Laden"*, *"Soviets"*, *"bombs"*, *"rebel camps"*, *"United States"*, *"holy warriors"*, *"Hekmatyar"*, ...), a screen which relates a topical highlight with a set of thread 'proxies'. (For optimal use of space, the symbol [] stands for the topic, in the list of topics representative of each story line.) Figure 6 shows a summary, corresponding to the first thread.

This is for illustrative purposes only. No attempt is made here to indicate (by some kind of highlighting or similar typographcal convention) the topic stamps from whose perspective this summary is presented. Traces of ellided material are heavily marked; in a functional system these will allow further exploration into a richer summary, or even the full text of the document. As a side effect of the cohesion analysis (Section 3.2), we are able to substitute dangling pronouns with their full antecedents (as seen in the opening of the third paragraph).

Note the (obligatory) contraction of the title, and the override of the browser's rendering of <p> tag (on a screen with less than a dozen lines of text, blank lines for paragraph boundaries are profligate). Also, by default the summary is displayed with maximum compression; this satisfies the pri-

mary requirement of this 'genre' of news delivery, namely rapidly getting a good sense of what a particular sub-story thread is about; this is in fact what, at the outset, was made difficult by the need for repeated and prolonged scrolling. It is easier to scan through a couple of screenfuls, and request more information; than to browse in the opposite direction, and get overloaded. This summary fits in three 'pages', a dramatic compression from the original size of the document, while still retaining informative value and a sense of cohesive unity.

5 Conclusion

Hand-helds with wireless connections are becoming ubiquitous. 'Just in time' content delivery, combined with the opportunistic, serendipitous access to news sources, defines a new genre of on-line news. This is a natural extension of current Web portals, but instead of designing for a standard size desktop screen, this genre will take account of the individual and personal nature of hand-helds.

Existing technologies for news content analysis are not optimally suited to meet the constraints of hand-held displays. We have argued that the specifics of the task of information seeking on hand-helds are best met by special purpose methods of news analysis, carried out by dedicated processing, rather than by remote proxy servers. In this paper we discuss aspects of such demand-driven, context-sensitive, user-defined content abstraction framework; we outline a mix of text processing technologies which derive a rich summary base; and we show how such a summary can be customised to reflect the interests of the current reader.

Our cohesion-based approach has been evaluated against an existing summariser, with its own special purpose evaluation test-bed using a corpus of full-length *New York Times* articles and their 'digests'. A number of experiments compare a base summarisation procedure, which calculates object, and based on that, sentence salience, with enhanced procedures incorporating different cohesion analysis strategies using notions of simple lexical chaining and topicality. Overall, leveraging cohesion analysis into the summarisation process is positively beneficial; the effects are particularly strong where short summaries are required (Boguraev and Neff, 2000). At present, we have no data for evaluating sentence reduction; also, no user studies have been carried out, as the interface is still under development (Section 4).

Our framework uses the highly personal nature of hand-helds by assuming selective filtering of news and by adapting the presentation of content highlights to incorporate awareness of particular user interests. Additionally, in communicating the document abstractions by means of suitably

Figure 6: A summary for one topical thread.

framing topical highlights, the presentation of the summary takes into account the constraints of the medium.

References

Saliha Azzam, Kevin Humphreys, and Robert Gaizauskas. 1999. Using coreference chains for text summarization. In *Proceedings of the* ACL-99 *Workshop on Coreference and Its Applications*, Baltimore, MD.

Brack Baldwin and Thomas S. Morton. 1998. Dynamic coreference-based summarization. In *Proceedings of Conference on Empirical Models in Natural Language Processing (*EMNLP-98*)*, Granada, Spain.

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.

Rachel K.E. Bellamy, Wendy A. Kellogg, John T. Richards, and Calvin B. Swart. 2000. Exploring a new paradigm in e-groceries. In *Proceedings of IBM Ease of Use Conference*, Yorktown Heights, NY, May.

Branimir Boguraev and Christopher Kennedy. 1999. Salience-based content characterisation of text documents. In Inderjeet Mani and Mark T. Maybury, editors, *Advances in Automatic Text Summarization*. MIT Press, Cambridge, Mass.

Branimir Boguraev and Mary Neff. 2000. The effects of analysing cohesion on document summarisation. In *Proceedings of the 18th International Conference on Computational Linguistics,* COLING-2000, Saarbruecken, Germany, August.

Branimir Boguraev, 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.

Branimir Boguraev. 2000. Towards finite-state analysis of lexical cohesion. In *Proceedings of the 3rd International Conference on Finite-State Methods for NLP,* INTEX-3, Liege, Belgium, June.

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

Gregory Grefenstette. 1998. Producing intelligent telegraphic text reduction to provide an audio scanning service for the blind. In *AAAI Spring Symposium on Intelligent Text Summarization*, pages 111–117, Stanford, CA.

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

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

Hongyan Jing. 2000. Sentence reduction for automatic text summarization. In *Proceedings of the 6th Applied Natural Language Processing Conference*, Seattle, WA, May.

Kevin Knight and Daniel Marcu. 2000. Statistics-based summarization — step one: sentence compression. In *Proceedings of Conference of the American Association for Artificial Intelligence (*AAAI-2000*)*, pages 703–710.

Inderjeet Mani and Eric Bloedorn. 1999. Summarizing similarities and differences among related documents. *Information Retrieval*, 1(1):35–67.

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

R. Mohan, J.R. Smith, and C.-S. Li. 1999. Adapting multimedia Internet content for universal access. In *Proceedings of IEEE Transactions on Multimedia*, volume 1, March.

M. Okumura, H. Mochizuki, and H. Nanba. 1999. Query-biased summarization based on lexical chaining. In *Proceedings of* PACLING-99 *(Pacific Association for Computational Linguistics)*, pages 324–334, Waterloo, Ontario.

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

Peter Pirolli and Stewart Card. 1998. Information foraging models of browsers for very large document spaces. In *Proceedings of the Advanced Visual Interfaces Workshop (*AVI'98*)*, pages 83–93, Aquila, Italy. Association for Computing Machinery.

Qualcomm Incorporated, San Diego, CA, 1999. *pdQsuite (*TM*) User Guide, with Eudora and pdQbrowser*. http://www.qualcomm.com/pqqsuite.

H. Gregory Silber and Kathleen F. McCoy. 2000. Efficient text summarization using lexical chains. In *Procedings of the* ACM *Conference on Intelligent User Interfaces (*IUI-2000*)*, New York, NY. Association for Computing Machinery.

J.R. Smith, R. Mohan, and C.-S. Li. 1998. Transcoding Internet content for heterogenous client devices. In *Proceedings of IEEE ISCAS-98, Special Session on the Next Generation Internet*, September.

M.E. Soper. 1976. Characteristics and use of personal collections. *Library Quarterly*, 46:397–414.

Karen Spärck Jones. 1993. What might be in a summary? In Knorz, Krause, and Womser-Hacker, editors, *Information Retrieval 93: Von der Modellierung zur Anwendung*, pages 9–26, Universitatsverlag Konstanz.

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