Natural Language Engineering **1** (1): 1–21. Printed in the United Kingdom

Visualization-Enabled Multi-Document Summarization by Iterative Residual Rescaling

RIE ANDO, BRANIMIR BOGURAEV, ROY BYRD, MARY NEFF

IBM T.J. Watson Research Center, 19 Skyline Drive, Hawthorne, NY 10532, USA
E-mail: {rie1, bran, roybyrd, maryneff}@us.ibm.com

(*Received October 2002; revised October 2003*)

Abstract

This paper describes a novel approach to multi-document summarization, which explicitly addresses the problem of detecting, and retaining for the summary, multiple themes in document collections. We place equal emphasis on the processes of theme identification and theme presentation. For the former, we apply Iterative Residual Rescaling (IRR); for the latter, we argue for graphical display elements. IRR is an algorithm designed to account for correlations between words and to construct multi-dimensional topical space indicative of relationships among linguistic objects (documents, phrases, and sentences). Summaries are composed of objects with certain properties, derived by exploiting the many-to-many relationships in such a space. Given their inherent complexity, our multi-faceted summaries benefit from a visualization environment. We discuss some essential features of such an environment.

1 Motivation

This work focuses on the complementary questions of detecting multiple themes in document collections and presenting these to an end user. The research is driven by a few key observations.

For homogenous document collections, where it is reasonable to assume the prominence of a topic, multi-document summarization can be largely driven by this assumption. In contrast, for heterogenous collections, multi-document summarization needs to be sensitive to multiple topics. Consequently, we describe the use of a novel technology particularly well suited to this task, as it is driven by the notion of identifying themes representative of a document, and possibly running across documents. Mapping a document collection to a set of topics, however, makes it imperative that the end user is assisted in navigating and comprehending the resulting multi-threaded summaries.

1.1 Operational assumptions for multi-document summarization technology

The information seeking tasks that stand to benefit from a multi-document summarization (MDS) capability are numerous and diverse; still, it is possible to identify two broadly different contexts for its deployment.

On the one hand, specialized applications define the technological make-up of a range of systems which, for instance, focus on summarizing multiple news stories about *the same event* (Radev and McKeown, 1998; McKeown et al., 1999; Barzilay et al., 1999) or a *sequence of events* (Mani and Bloedorn, 1999), generating summaries of medical articles filtered from the perspective of *a specific patient's needs* (Elhadad and McKeown, 2001), producing *composite biographical 'sketches'* from information about people in dispersed news articles (Schiffman et al., 2001) and even synthesising *multimedia briefings* by gathering and processing domain- and genre-specific information from multiple sources (Mani et al., 2000), or summarizing spoken dialogue interactions (Reithinger et al., 2000; Zechner and Lavie, 2001). In general, a common feature to such work is the focused mining for, and subsequent fusion of, information snippets strongly 'aligned' by similar or related information about the same event or object across several sources.

A different view of the utility of MDS assumes that it operates as a post-processing component to an information retrieval system, typically a search engine responding to a particular query. The very existence of a *query* drives in a fundamental way the design of a multi-document summarizer, as the query defines the perspective from which information is judged for its relevance and appropriateness for inclusion in the summary. Details vary, as the query may be used to reflect the relevance of a document in the context of *user-focused summarization* (Mani and Bloedorn, 1998; Berger and Mittal, 2000), to assess *relevant novelty* in passage selection (Goldstein et al., 2000), or to be fused with a notion of *a user profile*, for improved usability (Radev et al., 2001).

The work we describe here takes a broader view on the operational contexts in which MDS might be effectively deployed. Both positions above share an underlying assumption of extrinsic homogeneity of the set of documents to be summarized, leading to a definition of summarization succinctly expressed by (Radev et al., 2001) as "selecting the most salient information in one or more textual documents". We seek to relax this: we recognize the existence of situations where a more heterogeneous set of documents needs to be processed; situations which cannot be adequately addressed by highlighting just one (or several, closely related) prominent topic(s). We also recognize the concomitant need of end users to be made aware of more complex 'information landscapes', arising from viewing MDS as a multiple-themes-aware function.

More specifically, our work addresses two complementary considerations. We regard the ability to respond to multiple themes in a document collection as crucial to a general purpose multi-document summarization capability; we also focus on the additional graphical components at the interface between the end user and the

MDS technology, needed to adequately convey the thematic landscape derived by the application of such a technology.

1.2 Visual representation and navigation within information spaces

Graphical overviews of large document landscapes[1] on the whole do not explicitly represent topicality; it is not, in fact, clear that they could do so. In general, the approach taken there is to consider documents as objects in high-dimensional space, and provide 2-D or 3-D representation of such document space. Examples of studies that address graphical presentation of multi-document spaces include the VIBE system (Olsen et al., 1993; Korfhage and Olsen, 1995), Galaxy (Rennison, 1994), SPIRE Themescapes (Wise et al., 1995), LyberWorld (Hemmje et al., 1994), and applications of self-organizing map utilizing neural network technique (Kohonen, 1997; Lin, 1993; Lagus et al., 1996). (Hearst, 1999) offers an excellent overview of document collection visualization issues.

Hearst also provides, elsewhere (Hearst, 1995), an example of effective use of topics in mediating a document space; the TILEBARS interface, however, is explicitly designed to show the proportional representation of query search terms in the resulting document hit-list, and it is not clear that its presentational features would generalize to larger, heterogeneous document collections being serendipitously browsed. For such browsing, methods like SCATTER/GATHER (Cutting et al., 1992) appear to be better suited. Still, such methods employ a very lightweight notion of topicality, with themes derived, indirectly, via unsupervised clustering. It has been observed (Hearst and Pedersen, 1996) that themes thus obtained differ among themselves in levels of description, and thus hold potential for confusing the user.

In any case, work on visualizing and navigating multi-document collections to date does not typically make provisions for, or use, multi-document summarization. Conversely, MDS efforts largely deliver their results in textual form. It is not clear that these two different perspectives on document collection analysis—multi-document summarization and document collection visualization, as they exist to date—are easy to combine, given the methodological differences in underlying technologies, and in particular those for document set modeling, similarity measures calculation, and document content proxy determination.

1.3 MDS as navigation through multiple topic space

In essence, we view the primary task of multi-document summarization to be that of identifying salient themes persistent across subsets of documents in the collection. In particular, this paper focuses on the synergistic deployment of *theme identification* and *theme presentation*. Even if we are unable to 'embody' a theme in

[1] The notion of "graphical overviews" should not be confused with graph-based summarization techniques, like those developed by e.g. (Salton et al., 1997; Mani and Bloedorn, 1997).

coherently generated prose, we start with the assumption that a mapping exists between a theme and a tightly connected (and threfore intuitively interpretable) set of coherent linguistic objects—such as phrases and sentences—which would act as a 'prompting' device when presented to the user in an appropriate context. As will become clear in the rest of the paper, we refer to such themes as *topics*.

We will further argue that an approach to topic-based multi-document summarization needs to incorporate techniques for visualization of document collections in ways which facilitate mediating the rich and complex, many-to-many, relationships between individual documents and linguistic objects serving as *topic descriptors*.

Iterative Residual Rescaling (IRR) is an independently developed method (Ando, 2000; Ando and Lee, 2001), which constructs a vector space indicative of relationships among documents, topical phrases, and sentences. The properties of such a space (see Section 2 below) facilitate the identification of those particular linguistic objects which are strongly associated with underlying themes in the collection; this makes IRR especially appropriate as a background technology for our MDS.

In the following two sections we focus on the adaptation and application of Iterative Residual Rescaling as an algorithmic procedure underlying MDS. Detailed discussion of the design and implementation of an interface to such a functionality is outside of the scope of this paper; however, given the emphasis we place on the synergy between technology and presentation, we also outline, in the penultimate section, a number of essential features of an interface minimally required to visualize the multi-faceted relationships among topics, topic descriptors, and documents which are at the heart of our approach.

2 Iterative Residual Rescaling for topic space construction

The technology underlying our framework relies on the construction of a vector space with certain properties, which we call a *topic space*. The notion is to map all the linguistic objects in the document collection to vectors in a space where directional closeness between vectors serves as a measure of topical similarity between corresponding linguistic objects. For our MDS strategy, we need a representational space to be capable of adequately capturing both complete documents and terse linguistic objects (such as phrasal units), whose granularity makes them suitable as topic descriptors.

A classical mapping method, the *Vector Space Model* (VSM) (Salton and McGill, 1983), which measures similarity essentially by term matching, is not adequate under our definition of MDS, since we need to assess topical proximity among objects smaller than complete documents—such as phrasal units and sentences—which may not share identical terms even if they happened to be topically related.

In this work, we adopt an independently developed subspace projection-based method, *Iterative Residual Rescaling* for constructing a topic space. IRR is a generalization of *Latent Semantic Indexing* (Deerwester et al., 1990) (LSI), and its advantages over LSI are empirically shown in (Ando and Lee, 2001). Precise descriptions of the IRR algorithm and its mathematical relation to LSI may be found

elsewhere (Ando, 2000). Here we only describe IRR's properties relevant to our multi-document summarization task.

IRR is an algorithm designed for generating vector representations for linguistic objects. First, it constructs a *subspace* on the basis of the *term* occurrence statistics observed in the documents. A subspace is a subset of a vector space—in this case, a subset of a conventional VSM space which is spanned by terms. Terms are typically 'content words' serving as constituents of other linguistic objects. Roughly speaking, IRR computes a subspace in such a way that it is spanned by linear combinations of frequently co-occurring terms, instead of individual terms. Linguistic objects are mapped to vectors by projecting their VSM vectors onto this subspace. Importantly, projection onto the subspace computed in this manner brings term vectors for frequently co-occurring terms close to each other. Furthermore, it accounts for term co-occurrences in a transitive manner—i.e., it brings vectors for terms co-occurring with the same term(s) close to each other. This is in contrast with VSM where term vectors are always orthogonal by construction—indicating zero topical similarities among any pair of terms.

As an example, when IRR is used to compute a subspace from documents in the computing domain, the vectors for *"keyboard"* and *"mouse"* would be close to each other, whereas the *"keyboard"* vector would be closer to that for *"guitar"* in the music domain. Consequently, phrase (or document) vectors—each of which is a weighted sum of vectors for its constituent terms—become closer to each other if their constituent terms frequently co-occur in the documents from which the subspace is computed. As we assume that term occurrences and topics are correlated, it follows that directional closeness of IRR vectors is a good indicator of topical similarity.

Since closeness of vectors corresponds to topical similarity in our topic space, topically similar documents should constitute a natural cluster. On this basis, we apply a clustering algorithm to document vectors in an IRR subspace to cluster topically close documents together. (Ando and Lee, 2001) have studied the performance of topical document clustering using IRR document vector representation. The study conducts a series of experiments, which apply several standard clustering algorithms to document vectors created by IRR, and evaluate how well proposed partitions match with the topic-based partitions where topics are judged by humans. IRR consistently outperforms LSI and VSM in a variety of settings. Ando and Lee also show that IRR facilitates a method of training for the number of clusters, which determines the number of topics to be detected. We note that any distance-based clustering method should produce desired clusters, if measurement of topical closeness is accurate, and if the number of clusters can be determined appropriately. In this work, we seek to gain leverage from IRR, which, by design, meets these two criteria.

After constructing document clusters indicative of topics in the document collection, we then compute an IRR subspace for each cluster, capturing in it term co-occurrence statistics specific to the topic represented by that cluster. We create a *topic vector* for each cluster so that it encapsulates the predominant concept in such a topic-focused space. Note that topic vectors are abstractions over constella-

tions of linguistic objects, but not linguistic objects themselves (see the more formal definition of a topic vector below, Section 3.1.2, p. 9). At best, it is collections of linguistic objects at different levels of granularity (terms, phrases, sentences), and their relationships to the topically prominent documents, that represent different perspectives on the topic.

Under such a definition of topic vectors, we can then measure topical proximity between dominant concepts and linguistic objects by comparing topic vectors with linguistic object vectors. As we elaborate below (Section 3.1.2), a subset of linguistic objects with the strongest association with the topic vector adequately represent that topic (see the opening discussion in Section 4, and Appendix, for examples).

Conceptually, such a subset would then act as a summary of the topic. Thus, one can imagine a process of taking the sentences computed as closest to the topic and presenting them as the summary. Note, however, that there would be—by definition—more than one topic in the document collection, and hence more than one summary. And, while we might reasonably expect that any sentence from any document might be used in one such summary only,[2] there is nothing to prevent—indeed, it is a feature of our method for computing, and using, topic spaces—linguistic objects of smaller granularity, namely phrases, from appearing in more than one of the topic-defining sets as described above. Thus, if we want to fully utilize the richness of our notion of topic space, and in particular the spaces underlying the topics in the collection, we need to be able to mediate the many-to-many relationships between the full complement of linguistic objects, topics, and documents. We elaborate on some of the related issues in Section 4 below, which looks at the presentational aspects of our strategy.

3 Extracting topical content

This section focuses on the technological aspects of our approach to MDS. We briefly outline below (Section 3.1) the essence and the use of the IRR algorithm, which is the essential component of our MDS method. Since prior studies by (Ando and Lee, 2001) have already analyzed the performance of IRR in comparison with other vector space techniques, and the performance of clustering methods over document vectors created by IRR, we focus here on a crucial question concerning the adequacy of the linguistic objects 'closest' to the topic vectors (hereafter, *topic descriptors*) to function as representative of topics. The second part of this section (3.2) thus describes experiments we conducted in order to evaluate the effects of noise on the selection of topic descriptors.

[2] There is no guarantee of this, however. For instance, news stories under different bylines might incorporate data from the same source(s), resulting in duplication at sentence level.

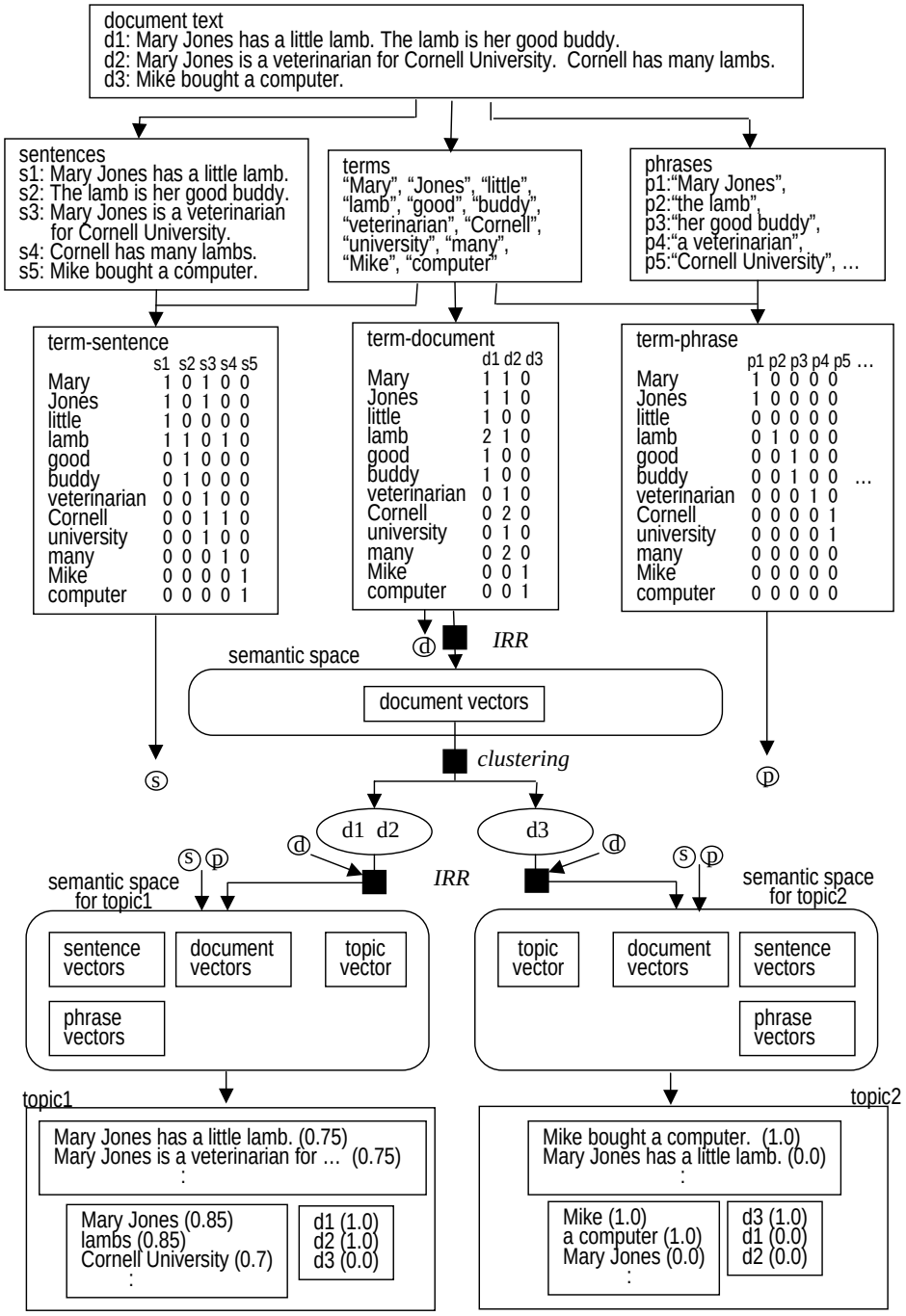

Fig. 1. Multi-document summarization by IRR: schematic illustration of analysis process.

3.1 Method

The method of data analysis underlying topical content identification and extraction is presented schematically in Figure 1. Overall, the data flow reflects the structuring of the MDS process in two stages: *identification of topics* and *selection of topic descriptors*. Both of these stages make use of IRR: initially, for the construction of the collection-level topic space, prior to clustering; and subsequently, for the secondary computation of topic (sub-)space within clusters, from which prominent topic descriptors are selected. Note that IRR is used in both cases to map sets of documents and linguistic objects to (different) topic spaces; and that it is the topic-preserving and -amplifying properties of this mapping that our MDS method crucially relies on. There is not, however—nor should there be—any relationship, assumed or expected, between the application of IRR in the first step, and that in the second. In particular (as we will see below in Section 3.1.2), the two invocations operate under different settings: this makes for the optimal (i.e. most resilient to noise) selection of topic descriptors.

Figure 1 traces, for the purposes of exemplifying a collection analysis, three 'mini-documents' (*d1*, *d2*, and *d3*), at the top of the diagram. While clearly a gross over-simplification, these are sufficient to illustrate the kind of linguistic objects (terms, phrases, and sentences) derived within a collection, and the data streams (matrices and vectors) that the algorithms traffic in. Omitting details of the particular processes for linguistic object extraction and data stream population, there are three pivotal points where other processes get invoked: the two stages of IRR invocation, and the clustering over the document vectors in the initial, collection-level, topic space. The end result—N topics, represented by prominent linguistic objects within the topics' respective topic spaces—is directly extractable from the vector representations after the secondary IRR application (note the coefficients for each linguistic object in the topic spaces, at the bottom of the figure: these impose a rank ordering with respect to the salience of an object relative to a given topic).

3.1.1 Topic identification

As already discussed, the initial application of IRR creates a collection-level topic space, whose properties reduce the problem of topic detection to one of clustering, so that documents related to a topic get grouped together. The input to IRR is a term-document matrix whose $[i, j]$-th entry represents the association between the i-th term and the j-th document. We apply the standard complete-link algorithm to IRR document vectors. (Manning and Schütze, 1999) give an excellent overview of clustering algorithms commonly used in natural language processing.

3.1.2 Topic descriptors selection

After clustering, each document cluster corresponds, roughly, to one topic. We say 'roughly' since, in practice, a machine-generated cluster may contain 'noise' documents (i.e. ones topically different from the dominant class of documents in that cluster). In particular, documents conveying multiple topics are destined to be

outliers regardless of which clusters they end up in. Our next task, then, is to create a topic vector to represent predominant concepts for each cluster while reducing the influence from such diffuse documents. Note that this has to be done *without* external knowledge of which documents should be regarded as noise.

To do this, we again rely on the properties of a topic space derived by IRR. We apply IRR to the documents in each cluster, and choose the first basis vector[3] of the resultant IRR subspace to be a *topic vector*. The choice is motivated by the fact that the first basis vector represents the most prominent term co-occurrence—and therefore the most dominant concept—in that cluster.

Thus, we use fundamentally the same IRR algorithm in two different settings: during the initial clustering, we seek to identify *all* topics in the collection; once clustering is done, we wish to ignore all but *the most dominant* topic in each cluster, thus obtaining more tight topical content. The effectiveness of this strategy for noise reduction will be experimentally confirmed in Section 3.2.

For each cluster, we measure the strength of associations between the topic and linguistic objects by comparing vectors in the IRR subspace (these are illustrated by the coefficients computed for the objects subsumed by topics in the topic diagrams, at the bottom of Figure 1). We choose k objects—phrases and sentences—having the strongest associations with that topic as topic descriptors. As with any MDS strategy which relies on re-using fragments and passages from documents, ours too is exposed to the problem of *redundancy*: duplicate, and/or very similar strings, and repeated substrings among linguistic objects, ultimately diffuse the information-bearing quality of a summary. Therefore, the selection of topic descriptors is further constrained by a redundancy filter, defined simply to eliminate from consideration objects sharing more than r% terms with the set of already chosen topic descriptors.

Just as training for the number of clusters and the dimensionality of IRR subspace (Section 2) is part of the overall system setup, the parameters k and r depend on the intended application. For instance, an interactive system might let users specify k; similarly, the value of r may depend on the presentation scheme of the resultant summary (see Section 4 below). In principle, however, any configured MDS application would require initial settings for these parameters.

Other types of associations, for instance those between documents and topic descriptors and those among topic descriptors, may be additionally useful when appropriately presented to users. In the IRR subspace, they can also be measured by comparing vectors: recall that one of the topic-preserving properties of IRR is that it allows for measuring associations between objects without term-sharing such as *"arms sales"* and *"weapons export"*. Prominent relationships of this nature can be usefully exploited, as MDS 'enhancers', via a suitable presentation component; we return to this in Section 4.8 below.

[3] *Basis vectors* are mutually orthogonal unit vectors that span a subspace. In the IRR subspace computation, the first basis vector is generated so that it maximally fits with the initial document vectors.

3.2 Experiments

The adequacy of linguistic objects 'closest' to a document cluster's topic vector to function as topic descriptors is clearly related to the ability of the method described above to deal with noise in the cluster. Here we describe experiments we conducted to measure the effectiveness of noise reduction in topic descriptor selection.

3.2.1 Data and metric

We used the TREC document collection to generate forty document sets, each one simulating one document cluster. The documents comprising the sets were taken to be relevant[4] to one of five (TREC) topics. We injected different levels of noise into the clusters, by having the sets have the following 2-, 3-, 4-, and 5-topic distributions, of ten sets for each: (45, 5), (40, 5, 5), (38, 4, 4, 4), and (38, 3, 3, 3, 3); in a distribution $(N_1, ..., N_n)$ (with n ranging from 2 to 5), N_i denotes the number of documents relevant to the i-th topic. We regard the documents on the topics of smaller populations as 'noise', e.g., a set of (38, 3, 3, 3, 3) contains $3 \times 4 = 12$ noise documents.

For each set, we selected the top k linguistic objects, where k ranges from 10 to 100 in increments of 10, and measured the overlap between our selection and the linguistic objects in documents on the dominant topic of the given set (3.2.2 below describes how the linguistic objects were extracted for this experiment). In case an object was found both in a dominant class of documents and in minority-topic documents, we associated the object with the topic in which it has larger relative occurrences. More formally, let X_k be the set of k top-ranked objects with respect to a topic vector. We define precision $p(X_k)$ as follows:

$$p(X_k) = \frac{|S \cap X_k|}{k}$$

where S is a set of linguistic objects associated with the dominant topic.

Arguably, the setup of our experiments does not truly capture real 'noise', and an ideal evaluation would be against human judgements concerning the degree to which each linguistic object in a document collection is either representative of a topic in that collection, or to be discounted as immaterial. Given the scope of an MDS task, such judgements would be required for the tens of thousands of linguistic objects typically comprising a document collection. This is clearly impractical; moreover, attempts to scale down the problem by appealing to subjective evaluation would raise the issue of inter-human agreement over what would still be a significant amount of data. We have therefore developed the statistical machinery described above as a reasonable approximation.

[4] TREC data documents are annotated with human judgements concerning their 'relevance' to a topic.

3.2.2 Implementation

In addition to sentences (and instead of terms only), the set of linguistic objects used to represent topical content includes full noun phrases; these were extracted from the documents by a shallow parser, functionally similar to the one described in (Boguraev and Kennedy, 1999). We removed phrases that contain pronoun(s) (e.g., *"his visit"*) as clearly being inappropriate for topic descriptors.[5]

To create a term-document matrix for input to IRR (see Section 3.1.1 and Figure 1), we used a standard $tf \times idf$ (Salton and McGill, 1983) and length-normalized document vectors.

Throughout the experiments, we observed that the values of parameter r (which controls the degree of term redundancy) do not affect the trend of relative performance with respect to baselines. For simplicity, we report the results of the most strict setting: with r being set to zero, which allows no term sharing with already-chosen objects.

3.2.3 Results

As a baseline (and baseline only) we created topic vectors by averaging the document vectors and measured associations by a simple cosine measure, without applying IRR.

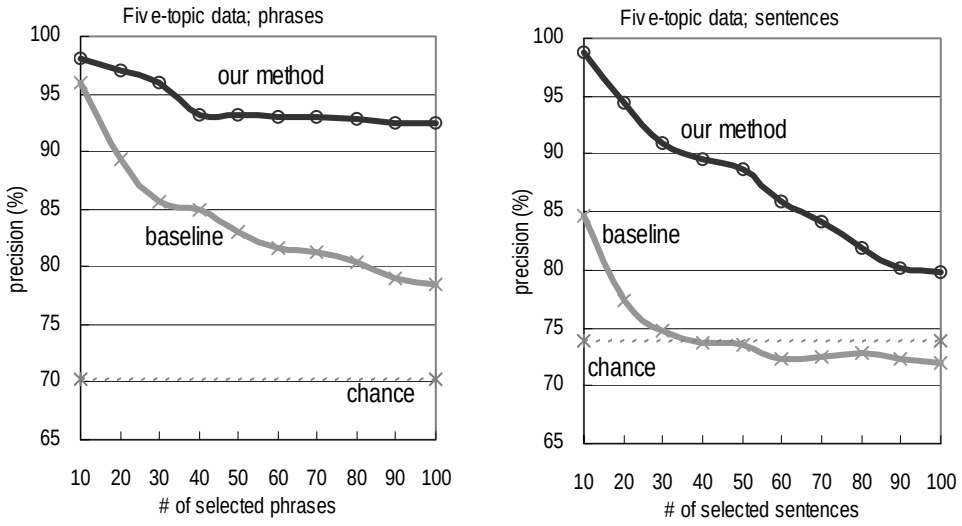

Fig. 2. Precision on five-topic data.

Figure 2 plots the precision (in the sense defined above) of topic descriptor se-

[5] In the absence of a coreference resolution capability.

lection on five-topic data averaged over ten sets. We analyzed performance for different linguistic object types—phrases and sentences—separately; thus two plots in the figure illustrate the precision of phrase, and sentence, selection. The x-axis is the number of selected phrases or sentences, as the parameter k grows. The horizontal dashed line shows the precision obtained by chance.

For the remainder of the data—four-, three- and two-topic distributions sets—the precision is tabulated in Figure 3. Our method of topic vector creation, and subsequent topic descriptors selection—produces higher precision than the baselines in all the settings: a result which clearly validates not only the selection procedure per se, but more importantly, its robustness against unavoidable noise in the topic clusters.

no. of selected phrases		10	50	100	no. of selected sentences		10	50	100
(45,5)	IRR-based	**98.0**	**97.4**	**96.3**	(45,5)	IRR-based	**98.0**	**91.3**	**86.6**
chance:82.6	baseline	97.0	93.3	90.5	chance:85.0	baseline	94.9	86.7	82.0
(40,5,5)	IRR-based	**97.0**	**94.1**	**91.7**	(40,5,5)	IRR-based	**97.0**	**87.2**	**85.7**
chance:78.4	baseline	94.8	83.4	80.8	chance:82.5	baseline	83.9	76.7	79.2
(38,4,4,4)	IRR-based	**96.0**	**92.1**	**87.3**	(38,4,4,4)	IRR-based	**93.9**	**83.3**	**78.2**
chance:70.3	baseline	88.0	79.6	77.7	chance:73.8	baseline	81.8	73.9	71.8

Fig. 3. Precision (%) on two-, three-, and four-topic data. The higher precision is highlighted.

4 Visual presentation of a topic space: combining text and graphics

This paper argues that our particular strategy for MDS places equal emphasis on the complementary processes of multiple topic *identification* and *presentation*. The interplay of topic descriptors of varying linguistic granularity with the rich topical content typically arising from retaining more than one salient topic for the purposes of a summary, crucially requires a mechanism for managing and mediating the relationships among topic descriptors and topics. Therefore, it is integral to our argument that for optimal appreciation of summaries which are not easy to couch in terms of linear text, a special purpose interface is required, which makes use of certain features (presentational metaphors) capable of visualizing the relationships which are intrinsic part of our summaries.

Complete user-centered design (Norman and Draper, 1986) and evaluation of a fully functional interface which accounts for end user information-seeking needs, would be the subject of a different paper. Interested readers are referred to (Boguraev et al., 1999) for an example of such design. Here, we are more concerned with analyzing the particular components of a visual environment as they derive from the properties of a topic space; we do this by showing a possible design for a sum-

mary layout, intended as an illustration of how textual and graphical elements could be combined.

The discussion below focuses on a summary for a document collection derived from TREC data. We have compiled a document collection by selecting 55 documents related to two (original) TREC topics (see footnote 4 earlier): *"non-proliferation treaty"* and *"firearms sales and crimes"*. (This ensures that the collection is 'tighter' than documents selected completely at random, an assumption we discuss earlier in Section 1.)

Our MDS method detected six topics; if we were to put concise labels on them, we might say that the topical clusters contain documents related, more specifically than TREC's topics, to *"Iraq"*, *"gun control"*, *"heavy water"*, and so forth, as illustrated in the Appendix. In general, the number of topics detected depends on the nature of the document set and on the parameter setting adjusting the granularity of analysis by IRR (see (Ando and Lee, 2001), and Section 2); we return to this below.

Figure 4 illustrates a visual layout of the topic descriptors, documents, and relationships among these, as they are 'read off' from the topic spaces. For the time being, we shall ignore the document proxy highlighting annotations (see note at

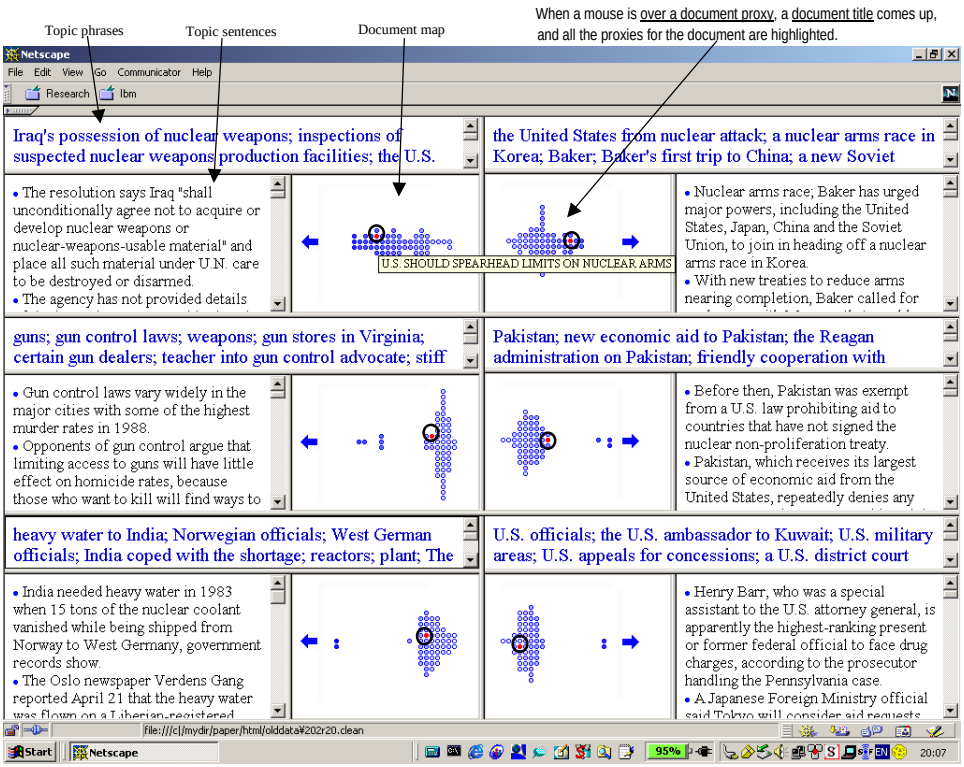

Fig. 4. Topical summary of a multi-document collection: screen display, and display dynamics via document proxies.

top right of the figure), which are not part of the display, and document title pop-ups, which need not be visible initially.

For each topic, three types of information are presented, grouped together: a list of phrasal units (*topic phrases*), a list of sentences (*topic sentences*), and a visual representation of the strength of the association between each document and the topic (*document map*); as indicated at the top left of the figure. As discussed above (3.1.2), topic phrases and sentences are selected *independently of each other*, hence the need for independent display areas for these. The document map is populated with data read out from the topic space, and offers a view on the document distribution ultimately correlated with the particular clustering results.

Below we highlight some of the essential features of such a display. Note that a finished design would, just as this prototype sketch does, seek effective use of color. In the description which follows, we approximate this by reference to shading and intensity.

4.1 Screen utilization

Given the complexity of topical information and variety in topical highlights, the available display area should be maximally used. The summary is thus presented in one full screen, in relation to the underlying topics. Naturally, provisions need to be made for the display of collections with different numbers of topics, e.g. by stacking screens to accommodate larger numbers.

4.2 Topic descriptors

Topic sentences can be used to present a conventional sentence-extraction based summary in a scrollable pane; topic phrases occupy a separate, independently controlled, pane. The joint display facilitates very efficient communication of the gist for a given topic; consider the combined effect of the topic phrases *"Iraq's possession of nuclear weapons"*, *"inspections of suspected nuclear weapons production facilities"*, and just the first of the topic sentences for topic 1 (top left corner of Figure 4), *"The resolution says Iraq 'shall unconditionally agree not to acquire or develop nuclear weapons . . . "*.

4.3 Document proxies

Together, a set of topic phrases and sentences describes a topic, i.e. one 'thread' discussed in possibly several documents. While topic descriptors are naturally represented as text fragments, documents in a document map are best represented by document proxies. In our example, round 'bullets' act as such proxies. The topic sentence pane also incorporates document proxies: there is one before each sentence, standing for the document containing that sentence. This enables dynamic linking (see below) between the document map and topic sentences panes.

4.4 Document maps; document proxy placement and intensity

In a document map, the horizontal placement of each proxy (with respect to the direction of an arrow) represents the degree of relatedness of the corresponding document to the topic. Documents closest to a topic sentence pane edge are central to the topic. The intensity of the proxy (progressive shading in the gray-scale figure, color gradation in reality) is also used to further convey the same information; see the document map for the *'Iraq'* topic, top left.

For instance, in the document map for the *'Pakistan'* topic (Figure 4, middle-right), three documents immediately stand out from the rest: they are visibly much closer to the topic sentence pane (the arrow in the document map pane marks an axis along which similarity among documents and the topic is to be interpreted); also, they are more intensely highlighted than other document proxies. The rest of the documents are not very related to this topic, which is indicated by the greater transparency of the proxies and the 'remoteness' of their placement. The intent is to enable users to tell, at a glance, how many documents are related to each topic and how strong that relationship is; the design exemplified here also allows comparative assessments to be made as to the configurations of topic clusters.

Note that unlike a typical presentation of document clustering results, we do not divide documents into groups: each document map contains proxies for all the documents; clusters of documents are naturally observed.

4.5 'Live' document proxies

The topic space allows us to detect and straightforwardly present the many-to-many relationships among documents and topics. When a mouse rolls over a document proxy, the title of the associated document is displayed, and the proxies for the same document in all the document maps are made to 'jump to attention' (shown by the highlighting circular marks in Figure 4). This is a different way of dynamically linking information in related panes, which facilitates understanding of the relationships between a document and multiple topics.

In this example, the document "U.S. SHOULD SPEARHEAD LIMITS ON NUCLEAR ARMS" can be seen to be related primarily to two topics: Iraq and its secret nuclear weapons programme (top left), and US' international efforts on nuclear non-proliferation (top right).

4.6 Mediating between text and graphics

By setting up document proxies as 'hot-links', clicking on a proxy would cause the full text of the corresponding document to be displayed in a separate window. This allows browsing of documents in the context of document-topic relationships.

4.7 Dynamic document annotation

A document opened via a proxy could be dynamically annotated for phrases associated with the particular topic in whose document map pane the interaction

was initiated. An instance of dynamic document annotation arises by clicking on a proxy associated with a topic sentence. In this case, the full text is displayed in a separate window, with the topic sentence highlighted. This highlighting facilitates understanding the context of the sentence quickly, and accelerates finding and focusing on information of particular interest.

As our method for topic descriptor selection (3.1.2) instantiates a topic space with measures of associations between linguistic objects and documents, a different kind of dynamic linking might explore these associations by e.g. displaying, for any given topic phrase, an ordered list of the documents (titles and proxies) which contextualize this phrase, in a manner similar to sentence highlighting.

4.8 Topic phrases

The default representation of topic phrases is as a sequence in a scrollable window. It is possible to imagine more engaging contexts, where additional organization is appropriate, for instance: grouping such phrases hierarchically, on the basis of syntactic or lexicographic regularities; or grouping them semantically, on the basis of strong associations—even if there is no term sharing—between them, measured in the focused topic space.

4.9 Topic sentences

Sentences are not particularly wieldy for direct manipulation. Nonetheless, the summary pane acts as a focus for sentence-directed operations. In addition to modulating the display by adjusting the allowed redundancy among sentences (see the next item below), we are experimenting with salience-focused sentence simplification (Boguraev et al., 2001), as an additional means of controlling sentence display.

For completeness, the full text of the top three sentences for three of the topic summary panes in Figure 4 is given in the Appendix.

4.10 Parameter settings

There are a few parameters for the algorithms underlying our MDS method, which directly affect the summaries.

In particular, these are related to topic granularity g (which is a parameter for IRR and indirectly determines the number of clusters in the topic detection process; see Section 2), population density of topic descriptors k, and redundancy filter r (Section 3.1.2). Generally, an operational system would be set up (typically, at installation time) with reasonable defaults; thus a user would not be expected (or required) to understand and adjust these. Thus, for the particular analysis of the document collection exemplified here, parameters such as g and r were determined based on the observation of a disjoint document set.

At the same time, exposing such parameters and allowing for their dynamic modification—e.g. by means of slider bars controlling, for instance, the number

of topic phrases and / or sentences extracted, or the degree of term overlap among summary sentences—would further facilitate interactive exploration of topical topic spaces, leading to even better understanding of summaries.

In general, the relationship between the specifics of the user task and interface characteristics is complex (Pirolli and Card, 1998). However, 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 (Soper, 1976). Thus, in designing a summarization interface to a document collection, 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 collection, as a whole, as well as of any individual document, in appropriate topical context.

In essence, a design for visualization should seek such economy of actions, at the same time facilitating quick appreciation of the contents of a document space by appealing to user intuitions, and conforming to established patterns of information seeking activities. This is subject to a different kind of analysis, itself an integral part of the process of user-centered design (Norman and Draper, 1986). We reiterate here that the discussion in this section does not present a complete system; this is not our intent. Rather, we argue the point that support of browsing through a document collection with easy switching between different views—topic highlights (phrases), topic summaries (sentences), full document texts, and inter-document relationships—is crucial for the kind of multi-perspective analysis assumed by our approach to multi-document summarization.

5 Conclusion and further work

This paper proposes a framework for multi-document summarization of heterogenous document collections, designed to leverage the technical strengths of a novel topic space building method, Iterative Residual Rescaling. The framework derives its capabilities in equal parts from the properties of IRR and from special-purpose graphical interface elements, as it presents a summary as a 'constellation' of topical highlights. On the basis of IRR analysis, our method seeks to detect multiple topics threading a given document collection, which are described by extracting related phrases and sentences from the document text. A visualization component, capable of mediating the many-to-many relationships underlying associations of phrases, sentences, topics, and documents, is essential for full appreciation of IRR-derived summaries.

It has been our intent to focus on the interplay of multiple topic detection and presentation, as we start this work from a different operational definition of multi-document summarization. Within such a definition, a summary is as good, and as robust, as the technologies underlying the topic detection are accurate and resilient to noise. Such technologies, and in particular IRR and clustering, are discussed and evaluated at length elsewhere. We have therefore focused on evaluating one particular aspect of the overall process, namely that of the effectiveness of our

topic descriptors selection procedure against noise after clustering. The results we report in Section 3.2.3 confirm the viability of our method.

When a technology is to be exposed to end users, established principles of user-centered interface design must be appealed to in order to 'wrap' the technology in a usable interface. Such a design is outside of the scope of this paper. We do argue, however, that our approach to MDS views an interface not as an add-on, but as an integral part of a summarization framework. Thus we lay out a number of design considerations which derive from the semantic properties of our summaries. The discussion in Section 4 should be viewed not as a definitive description of a complete system—to be validated and evaluated against usability, habitability, and satisfaction criteria—but as a context in which essential visual metaphors and features (such as multi-pane, multi-modal presentation of semantic associations; linking between different views; dynamic document annotation; and so forth) are exemplified.

Indeed, we have hinted at numerous possibilities for alternative and/or additional renderings of topic groups within the semantic space; typically these would be mediated by different granularity of document fragments, and exploiting the semantic associations in topic-focused semantic space (Section 4). We have done some work on determining the effects of analyzing linguistic objects (e.g. sentence- and clause-level phrasal units, and different semantic categories of phrasal types) and their grammatical function, specifically for the purpose of representing closely related topical documents to the user. We are already developing methods for more rigorous structuring of the set of prominent topic descriptors.

Outside of the scope of this paper remain a number of open questions. What kinds of phrases are adequate, and optimal, 'carriers' of topical content? How much would operations over sentences, such as sentence merging or simplification, offer to alternative ways of visualizing topical content? To what extent is sentence ordering, e.g. in the sense of (Barzilay et al., 2001), strictly necessary for an approach like ours, which assumes multiple views into the topics space? These are essential enabling technologies for any kind of robust summarization technique, and the exact strategies for tackling them within our IRR-based framework is the subject of future research.

Acknowledgements

We thank Herb Chong, James Cooper, Lillian Lee, Alan Marwick, Tetsuya Nasukawa, John Prager, Dragomir Radev and Edward So for helpful discussions and for valuable comments and suggestions; we are especially indebted to detailed, anonymous, reviews on an earlier draft. The first author was partly supported by a McMullen fellowship from Cornell University.

References

Ando, R. K. (2000). Latent semantic space: Iterative scaling improves inter-document similarity measurement. In *Proceedings of SIGIR'2000*, pages 216–223.

Ando, R. K. and Lee, L. (2001). Iterative residual rescaling: An analysis and generalization of LSI. In *Proceedings of SIGIR'2001*.

Barzilay, R., Elhadad, N., and McKeown, K. R. (2001). Sentence ordering in mulidocument summarization. In *Proceedings of HLT-01*, San Diego, CA.

Barzilay, R., McKeown, K. R., and Elhadad, M. (1999). Information fusion in the context of multi-document summarization. In *Proceedings of the 37th Annual Meeting of the ACL*.

Berger, A. and Mittal, V. O. (2000). Query-relevant summarization using FAQs. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 294–301, Hong Kong.

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

Boguraev, B., Bellamy, R., and Swart, C. (2001). Summarisation miniaturisation: delivery of news to hand-helds. In *Proceedings of Workshop on Automatic Summarization, NAACL-2001*, pages 99–108, Pittsburgh, PA.

Boguraev, B. and Kennedy, C. (1999). Salience-based content characterisation of text documents. In Mani, I. and Maybury, M. T., editors, *Advances in Automatic Text Summarization*, pages 99–110. MIT Press, Cambridge, MA.

Cutting, D. R., Pedersen, J. O., Karger, D., and Tukey, J. W. (1992). Scatter / Gather: a cluster-based approach to browsing large document collections. In *Proceedings of the 15th Annual International ACM/SIGIR Conference*, pages 318–329, Copenhagen, Denmark.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the Society for Information Science*, 41:391–407.

Elhadad, N. and McKeown, K. R. (2001). Towards generating patient specific summaries of medical articles. In *Proceedings of Workshop on Automatic Summarization, NAACL-2001*, pages 32–40, Pittsburgh, PA.

Goldstein, J., Mittal, V., Carbonell, J., and Kantrowitz, M. (2000). Multi-document summarization by sentence extacton. In *Proceedings of Workshop on Automatic Summarization, ANLP/NAACL-2000*, pages 40–48, Seattle, WA.

Hearst, M. A. (1995). Tilebars: Visualization of term distribution information in full text information access. In *ACM SIGCHI Conference on Human Factors in Computing Systems*, pages 59–66, Denver, CO.

Hearst, M. A. (1999). User interfaces and visualisation. In Baeza-Yates, R. and Ribeiro-Neto, B., editors, *Modern information retrieval*, pages 257–324. Addison-Wesley, New York.

Hearst, M. A. and Pedersen, J. O. (1996). Reexamining the cluster hypothesis: scatter / gather on retrieval results. In *Proceedings of the 19th Annual International ACM/SIGIR Conference*, pages 76–84, Zurich, Switzerland.

Hemmje, M., Kunkel, C., and Willett, A. (1994). LyberWorld – a visualization user interface supporting fulltext retrieval. In *Proceedings of SIGIR'94*, pages 249–260.

Kohonen, T. (1997). Exploration of large document collections by self-organizing maps. In *Proceedings of SCAI'97*, pages 5–7.

Korfhage, R. R. and Olsen, K. A. (1995). Image organization using vibe a visual information browsing environment. In *Proceedings of SPIE*, volume 2606, pages 380–388.

Lagus, K., Honkela, T., Kaski, S., and Kohonen, T. (1996). Self-organizing maps of document collections: A new approach to interactive exploration. In *Proceedings of Second International Conference on Knowledge Discovery & Data Mining*, pages 238–243.

Lin, X. (1993). Map displays for information retrieval. *Information Processing & Management*, 29(1):69–81.

Mani, I. and Bloedorn, E. (1997). Multi-document summarization by graph search and matching. In *Proceedings of AAAI'97*, pages 622–628.

Mani, I. and Bloedorn, E. (1998). Machine learning of generic and user-focused summarization. In *Proceedings of AAAI-98*, pages 821–826, Madison, WI.

Mani, I. and Bloedorn, E. (1999). Summarizing similarities and differences among related documents. *Information Retrieval*, 1(1):1–23.

Mani, I., Concepcion, K., and Guilder, L. V. (2000). Using summarization for automatic briefing generation. In *Proceedings of Workshop on Automatic Summarization, ANLP/NAACL-2000*, pages 99–108, Seattle, WA.

Manning, C. D. and Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press, Cambridge, MA.

McKeown, K. R., Klavans, J., Hatzivassiloglou, V., Barzilay, R., and Eskin, E. (1999). Towards multidocument summarization by reformulation: Progress and prospects. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI-99)*.

Norman, D. A. and Draper, S. W. (1986). *User-centered system design: new perspectives on human-computer interaction*. LAwrence Erlbaum Associates.

Olsen, K. A., Korfhage, R. R., Sochats, K. M., Spring, M. B., and Williams, J. G. (1993). Visualization of a document collection: The VIBE System. *Information Processing & Management*, 29(1):69–81.

Pirolli, P. and Card, S. (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.

Radev, D. R., Fan, W., and Zhang, Z. (2001). WebInEssence: a personalized web-based multidocuent summarizaton and recommndation system. In *Proceedings of Workshop on Automatic Summarization, NAACL-2001*, pages 79–88, Pitsburgh, PA.

Radev, D. R. and McKeown, K. R. (1998). Generating natural language summaries from multiple on-line sources. *Computational Linguistics*, 24(3):469–500.

Reithinger, N., Kipp, M., Engel, R., and Alexandersson, J. (2000). Summarizing multilingual spoken negotiation dialogues. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 310–317, Hong Kong.

Rennison, E. (1994). Galaxy of news: An approach to visualizing and understanding expansive news landscapes. *presented at the ACM Symposium on User Interface Software and Technology, Marina del Rey, CA, November*, pages 2–4.

Salton, G. and McGill, M. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill.

Salton, G., Singhal, A., Mitra, M., and Buckley, C. (1997). Automatic text structuring and summarization. *Information Processing and Management*, pages 193–207.

Schiffman, B., Mani, I., and Concepcion, K. J. (2001). Producing biographical summaries: combining linguistic knowledge with corpus statistics. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, pages 450–457, Toulouse, France.

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

Wise, J. A., Thomas, J., Pennock, K., Lantrip, D., Pottier, M., Schur, A., and Crow, V. (1995). Visualizing the non-visual: spatial analysis and interaction with information from text documents. In *Proceedings of Information Visualization*, pages 51–58.

Zechner, K. and Lavie, A. (2001). Increasing the coherence of spoken dialogue summaries by cross-speaker information linking. In *Proceedings of Workshop on Automatic Summarization, NAACL-2001*, pages 22–31, Pittsburgh, PA.

Appendix

The following extracts illustrate the highly focused nature of IRR-based summarization.

'Iraq-topic':

- *The resolution says Iraq "shall unconditionally agree not to acquire or develop nuclear weapons or nuclear-weapons-usable material" and place all such material under U.N. care to be destroyed or disarmed.*
- *The agency has not provided details of the letter, but sources said privately that Iraq denied it had a nuclear weapons program and said it has complied with the agency's inspection program for its reactors.*
- *Blix said he showed the council a videotape of Iraqi nuclear sites that showed how Iraq was trying to hide evidence that it had a program to create highly enriched uranium, which is needed for nuclear weapons.*

'Gun control-topic':

- *Gun control laws vary widely in the major cities with some of the highest murder rates in 1988.*
- *Opponents of gun control argue that limiting access to guns will have little effect on homicide rates, because those who want to kill will find ways to obtain guns or will use other weapons.*
- *A study concludes that Seattle 's weaker gun control laws probably explain why the risk of being killed with a handgun there is five times higher than in nearby Vancouver, British Columbia.*

'Heavy water-topic':

- *India needed heavy water in 1983 when 15 tons of the nuclear coolant vanished while being shipped from Norway to West Germany, government records show.*
- *The Oslo newspaper Verdens Gang reported April 21 that the heavy water was flown on a Liberian-registered West African Airlines plane via Basel, Switzerland, to Dubai, then probably to a nuclear reactor in Bombay, India.*
- *Another official at the Ministry of Science and Technology flatly denied that the missing Norwegian heavy water came to India.*

Both for 'Iraq-' and 'heavy water-topic', the second and third sentences offer additional facts related to the subject brought up in the first sentence. For 'gun control-topic', three items of statistics and opinions on one theme—correlation between gun control laws and murder rates—are shown; it is worth pointing out that "*murder rates*" in the first sentence is expressed differently as "*homicide rates*" in the second,, and "*the risk of being killed*" in the third.