

Saliency-Based Content Characterisation of Text Documents

Branimir Boguraev

IBM T.J. Watson Research Center
P.O. Box 704, Yorktown Heights, NY 10598
bkb@watson.ibm.com

Christopher Kennedy

Department of Linguistics, Northwestern University
2016 Sheridan Road, Evanston, IL 60208
kennedy@ling.nwu.edu

Abstract

Summarisation is poised to become a generally accepted solution to the larger problem of content analysis. We offer an alternative perspective on this problem, by tackling the complementary task of content characterisation; our motivation for doing so is to avoid some of the fundamental shortcomings of summarisation technologies today. Traditionally, the document summarisation task has been tackled either as a natural language processing problem, with an instantiated meaning template being rendered into coherent prose, or as a passage extraction problem, where certain fragments (typically sentences) of the source document are deemed to be highly representative of its content, and thus delivered as meaningful “approximations” of it. Balancing the conflicting requirements of depth and accuracy of a summary, on the one hand, and document and domain independence, on the other, has proven a very hard problem. This paper describes a novel approach to content characterisation of text documents. It is domain- and genre-independent, by virtue of not requiring an in-depth analysis of the full meaning. At the same time, it remains closer to the core meaning by choosing a different granularity of its representations (phrasal expressions rather than sentences or paragraphs), by exploiting a notion of discourse contiguity and coherence for the purposes of uniform coverage and context maintenance, and by utilising a strong linguistic notion of saliency, as a more appropriate and representative measure of a document’s “aboutness”.

1 Introduction

As the volume of document-based information online continues growing, so does the need for *any* kind of document abstraction mechanism. Consequently, summarisation has become one of the hottest topics in applied natural language processing, and some summarisation technologies are rapidly gaining deployment in real world situations.

1.1 Document summarisation and content characterisation

The wide acceptance of the term “summarisation” does not reflect the fact that several constraints apply to the commonly shared notion of a document “summary”. First, summaries are assumed to be, in essence, small documents (smaller, in any case, than the originals). A summary thus may vary in size, over a range from a single sentence to one or more paragraphs, but there is always the expectation that it will

be delivered to its intended users in the form of coherently readable prose. To a large extent, the shared (and for the most part unspoken) intuition is that summaries are to documents very much what abstracts are to full-length articles. This reflects another pervasive assumption: namely that there is a canonical, definitive (or at least optimal) summary for any document. As a result, work on summarisation technologies tends to proceed outside of any strong considerations of the operational environments where users may seek to deploy summarisation. Indirectly, this tendency reinforces the summary-as-an-abstract view which underlies most examples of the document summarisation paradigm today.

Largely, such observations are attributable to the globally pervasive fact that all of the current work in the field is carried out without reference to any theory of summarisation. We share cognitive intuitions about what a summarisation technology might strive to develop as operational machinery, but there is no grounding of such technology in a framework which has something concrete to say about what it is that defines a summary, nor how such a summary should be related to its full document source. For instance, agreement on the need for relevance measures for summarisation still leaves open the questions of how relevance ought to be computed, what units it should be computed over, and how it best might drive a generation process which tries to weave a coherent statement about the relevant highlights of a document. Currently, summarisation work is very much about the whole package, and hardly at all about individual pieces which underlie the construction of summaries.

Only recently have these views been challenged. The TIPSTER/SUMMAC evaluation conference (Def 1998) defined *several* different uses for summarisation, thus taking the first step to acknowledging that different information management tasks are likely to require different kinds of summary, even from the same document. While the conference itself still focused on seeking one particular type of summary as a baseline, in general, the community is becoming much more attuned to the fact that there is no such thing as a ‘canonical’ summary for a document; see e.g. (Sparck Jones 1997). However, there is much less discussion about the genre characteristics of the different summary types themselves.

There have been some implicit references to summary types departing from the notion of a small, coherent document. Library cataloging services, for instance, effectively utilise summaries in the form of key

index terms; technical documents of average-to-large size can be 'abstracted' by using a mix between table of contents and back-of-the-book index; cross-language 'gisting' using indiscriminating phrasal spotting can convey something of the content of a document to determine whether it should be fully translated or not (Endres-Niggemeyer 1998), (Resnik 1997). However, such work tends to be regarded as peripheral to mainstream summarisation research which focuses, primarily, on deriving document-like document abstractions.

This paper reports on some work which departs from the mainstream view, while still seeking to address the general problem of conveying to a user the gist of a document. Instead of focusing on summarisation, in its accepted sense, we define the problem to be that of characterising the essential content of a document. Again, in the absence of a strong theory, we appeal to intuitions concerning the relationship between the distributional prominence of linguistic expressions, computed as a function of their occurrence in a text, and the topical prominence of the objects and events they refer to. We seek to characterise a document's content by identifying a (relatively) brief enumeration of those expressions that refer to the most prominent, or most *salient* objects and events mentioned in the discourse.

These are similar intuitions to those shared by the summarisation community; the differences in our position lie in what we consider to be the linguistic units that should be targeted for the purpose of content abstraction, how we determine a measure of the relative prominence—or salience—of these units, how this measure is used to derive a document abstraction, and how the resulting abstractions are presented to users. The question of optimal presentation of our document abstractions for best use is outside of the scope of this paper; (Boguraev *et al.* 1998) discusses an experiment in situating our document abstraction technology in the context of supporting dynamic, on-line news skimming. Before we present details of our approach, and highlight the differences between content characterisation and document summarisation, we briefly outline the main characteristics of current approaches to summarisation.

1.2 Approaches to document summarisation

The majority of techniques for summarisation, as applied to average-length documents, fall within two broad categories: those that rely on template instantiation and those that rely on passage extraction. Work in the former framework traces its roots to some pioneering research by DeJong (1982) and Tait (1983); more recently, the DARPA-sponsored TIPSTER programme (Adv 1993b)—and, in particular, the message understanding conferences MUC: e.g. (Def 1992) and (Adv 1993a)—have provided fertile ground for such work, by placing the emphasis of document analysis to the identification and extraction of certain core entities and

facts in a document, which are packaged together in a template. There are shared intuitions among researchers that generation of smooth prose from this template would yield a summary of the document's core content; recent work, most notably by McKeown and colleagues, cf. (McKeown & Radev 1995), focuses on making these intuitions more concrete.

While providing a rich context for research in generation, this framework requires an analysis front end capable of instantiating a template to a suitable level of detail. Given the current state of the art in text analysis in general, and of semantic and discourse processing in particular, work on template-driven, knowledge-based summarisation to date is hardly domain- or genre-independent (see Sparck Jones 1993a, 1993b for discussion of the depth of understanding required for constructing true summaries).

The alternative framework—passage extraction—largely escapes this constraint, by viewing the task as one of identifying certain segments of text (typically sentences) which, by some metric, are deemed to be the most representative of the document's content. The technique dates back at least to the 50's (Luhn 1958), but it is relatively recently that these ideas have been filtered through research with strongly pragmatic constraints, for instance: what kinds of documents are optimally suited for being "abstracted" in such a way (e.g. Preston & Williams 1994; Brandow, Mitze, & Rau 1995); how to derive more representative scoring functions, e.g. for complex documents, such as multi-topic ones (Salton *et al.* 1996), or where training from professionally prepared abstracts is possible (Kupiec, Pedersen, & Chen 1995); what heuristics might be developed for improving readability and coherence of "narratives" made up of discontinuous source document chunks (Paice 1990); or with optimal presentations of such passage extracts, aimed at retaining some sense of larger and/or global context (Mahesh 1997).

The cost of avoiding the requirement for a language-aware front end is the complete lack of intelligence—or even context-awareness—at the back end: the validity, and utility, of sentence- or paragraph-sized extracts as representations for the document content is still an open question (see, for instance, Rau, 1988, and more recently, AAI 1998). This question is becoming more urgent, especially with the recent wave of commercial products announcing built-in "summarisation" (by extraction) features (Caruso 1997).¹ Nonetheless, progressively more sophisticated techniques are being deployed in attempts to improve the quality of sentence-based summaries, by seeking to mediate the passage selection process with, for instance, strong notions of topicality (Hovy & Lin 1997), lexical chains (Barzilay & Elhadad 1997), and discourse structure (Marcu 1997), (Reimer & Hahn 1997).

¹Also at: <http://www.nytimes.com/library/cyber/digicom/012797digicom.html>.

1.3 Capsule overviews

The approach we take in this work, while addressing a slightly different problem to that of strict summarisation, can be construed as striving for the best of both worlds. We use linguistically-intensive techniques to identify those phrasal units across the entire span of the document that best function as representative highlights of the document's content. The set of such phrasal units, which we refer to as *topic stamps*, presented in ways which both retain local and reflect global context, is what we call a *capsule overview* of the document.

A capsule overview is not a conventional summary, in that it does not attempt to convey document content as a sequence of sentences. It is, however, a semi-formal (normalised) representation of the document, derived after a process of data reduction over the original text. Indeed, by adopting finer granularity of representation (below that of sentence), we consciously trade in "readability" (or narrative coherence) for tracking of detail. In particular, we seek to characterise a document's content in a way which is representative of the full flow of the narrative; this contrasts with passage extraction methods, which typically highlight only certain fragments (an unavoidable consequence of the compromises necessary when the passages are sentence-sized). While we acknowledge that a list of topic stamps by itself is lacking in many respects as a coherent summary, we will argue that such a list is nevertheless highly representative of what a document is about, and, when combined with contextual cues associated with the topic stamps as they appear in the text, provides the basis for a useful, and informative, abstraction of document content.

Still, a capsule overview is—by design—not intended to be read the same way in which a document, or an abstract, would be. This paper focuses on the linguistic processes underlying the automatic identification and extraction of topic stamps and their organisation within capsule overviews. As already mentioned, the issues of the right presentation metaphor and operational environment(s) for use of topic stamps-based capsule overview are the subject of a different discussion; see, for instance, (Boguraev *et al.* 1998).

As this suggests, a capsule overview is not a fully instantiated meaning template. A primary consideration in our work is that content characterisation methods apply to any document source or type. This emphasis on *domain independence* translates into a processing model which stops short of a fully instantiated semantic representation. Similarly, the requirement for efficient, and scalable, technology necessitates operating from a shallow syntactic base; thus our procedures are designed to circumvent the need for a comprehensive parsing engine. Not having to rely upon a parsing component to deliver in-depth, full, syntactic analysis of text makes it possible to generate capsule overviews for a variety of documents, up to and including real data from unfamiliar domains or novel genres.

In its most basic respects, then, a capsule overview is composed of a list of the linguistic expressions referring to the most prominent objects mentioned in the discourse—the topic stamps—and a specification of the relational contexts (e.g. verb phrases, minimal clauses) in which these expressions appear. The intuitions underlying our approach can be illustrated with the following news article:²

PRIEST IS CHARGED WITH POPE ATTACK

A Spanish Priest was charged here today with attempting to murder the Pope. *Juan Fernandez Krohn*, aged 32, was arrested after *a man armed with a bayonet* approached the Pope while he was saying prayers at Fatima on Wednesday night.

According to the police, *Fernandez* told the investigators today that *he* trained for the past six months for the assault. *He* was alleged to have claimed the Pope 'looked furious' on hearing *the priest's* criticism of his handling of the church's affairs. If found guilty, *the Spaniard* faces a prison sentence of 15–20 years.

There are a number of reasons why the title, '*Priest Is Charged with Pope Attack*', is a highly representative abstraction of the content of the passage. It encapsulates the essence of what the story is about: there are two actors, identified by their most prominent characteristics; one of them has been attacked by the other; the perpetrator has been charged; there is an implication of malice to the act. The title brings the complete set of salient facts together, in a thoughtfully composed statement, designed to be brief yet informative. Whether a present day natural language analysis program can derive—without being primed of a domain and genre—the information required to generate such a summary is arguable. (This is assuming, of course, that generation techniques could, in their own right, do the planning and delivery of such a concise and information-packed message.) However, part of the task of delivering accurate content characterisation is being able to identify the components of this abstraction (e.g., '*priest*', '*pope attack*', '*charged with*'). It is from these components that, eventually, a message template would begin to be constructed.

It is also precisely these components, viewed as phrasal units with certain discourse properties, that a capsule overview should present as a characterisation of the content of a text document. Abstracting from the technicalities of automatic document analysis, the difference between a summary and a capsule overview thus could be informally illustrated by the difference between a summary-like statement, such as '*A Spanish priest is charged after an unsuccessful murder attempt on the Pope*', and an enumeration of the salient document highlights of the document, such as:

²Adapted from an example of S. Nirenburg; italics are ours, and are explained in Section 4.1 below.

A SPANISH PRIEST *was charged*
attempting to murder the POPE
HE *trained for the assault*
POPE *furious on hearing* PRIEST'S *criticisms*

Our strategy therefore is to mine a document for the phrasal units that are most representative of its content, as well as the relational expressions they are associated with, with the goal of establishing the kind of content characterisation exemplified here. The goal of this paper is to describe a procedure that implements this selective mining of a document for its most representative—its most *salient*—phrases, which we refer to as *salience based content characterisation*.

The remainder of this paper is organised as follows. Given the importance we assign to phrasal analysis, we outline in Section 2 and Section 3 the starting point for this work: research on terminology identification and the extension of this technology to non-technical domains. In particular, we focus on the problems that base-line terminology identification encounters when applied to open-ended range of text documents, and outline a set of extensions required for adapting it to the goal of core content identification. These boil down to formalising and implementing an operational, computable notion of salience which can be used to impose an ordering on phrasal units according to the topical prominence of the objects they refer to; this is discussed in Section 4. Section 5 illustrates the processes involved in topic identification and construction of capsule overviews by example. We close by positioning this work within the space of summarisation techniques.

2 Technical terminology: strengths and limitations

The identification and extraction of technical terminology is, arguably, one of the better understood and most robust NLP technologies within the current state of the art of phrasal analysis. What is particularly interesting for us is the fact that the linguistic properties of technical terms support the definition of computational procedures for term identification that maintain their quality regardless of document domain and type. What is even more interesting is that there is strong empirical evidence in support of the intuition that, in technical prose at least, terminological noun phrases are topical (Justeson & Katz 1995).

Since topic stamps as defined above are just phrasal units with certain discourse properties (they are topically prominent within contiguous discourse segments), we can define the task of content characterisation as one of identifying phrasal units that have lexico-syntactic properties similar to those of technical terms and discourse properties that signify their status as most prominent. In Section 4, we show how salience is computable as a function of the grammatical distribution of the phrase. Before moving to this discussion, however, we address the issues that arise when termi-

nology identification is applied to the content characterisation task.

One of the best defined procedures for technical terminology identification is the TERMS algorithm developed by Justeson & Katz (1995), which focuses on multi-word noun phrases occurring in continuous texts. A study of the linguistic properties of these constituents—preferred phrase structures, behaviour towards lexicalisation, contraction patterns, and certain discourse properties—leads to the formulation of a robust and domain-independent algorithm for term identification. Justeson and Katz's TERMS algorithm accomplishes high levels of coverage, it can be implemented within a range of underlying NLP technologies (e.g.: morphologically enhanced lexical look-up, Justeson & Katz, 1995; part-of-speech tagging, Dagan & Church, 1995; or syntactic parsing, McCord 1990), and it has strong cross-linguistic application (see, for instance, (Bourigault 1992)). Most importantly for our purposes, the algorithm is particularly useful for generating a “first cut” towards a broad characterisation of the content of the document.

Conventional uses of technical terminology are most commonly identified with text indexing, computational lexicology, and machine-assisted translation. Less common is the use of technical terms as a representation of the topical content of a document. This is to a large extent an artifact of the accepted view—at least in an information retrieval context—which stipulates that terms of interest are the ones that *distinguish* documents from each other. Almost by definition, these are not necessarily the terms which are representative of the “aboutness” of a document, as the expressions that provide important information about a document's content often do not distinguish that document from other texts within the same domain.

Still, it is clear that a program like TERMS is a good starting point for distilling representative lists. For example, (Justeson & Katz 1995, appendix) presents several term sets that clearly identify the technical domain to which the documents they originate in belong: ‘*stochastic neural net*’, ‘*joint distribution*’, ‘*feature vector*’, ‘*covariance matrix*’, ‘*training algorithm*’, and so forth, accurately characterise a document as belonging to the statistical pattern classification domain; ‘*word sense*’, ‘*lexical knowledge*’, ‘*lexical ambiguity resolution*’, ‘*word meaning*’, ‘*semantic interpretation*’, ‘*syntactic realization*’, and so forth assign, equally reliably, a document to the lexical semantics domain.

However, although such lists are representative, their size can easily become overwhelming. Conventionally, volume is controlled by promoting terms with higher frequencies. This is a very weak metric for our purposes, however, as it does not scale down well for texts that are smaller than typical instances of technical prose or scientific articles—such as news stories, press releases, or web pages. More generally, without the closed nature of technical domains and documentation, it is not clear that sets of term-like phrases de-

rived from arbitrary texts can provide the same level of informativeness as term sets derived from technical documents. Certainly, we cannot even talk of “technical terms” in the narrower sense assumed by the TERMS algorithm. This raises the following question: can the notion of technical term be appropriately extended, so that it applies not just to scientific prose, but to an open-ended set of document types and genres? In other words, can a set of phrases derived in this way provide a representational base which enables rapid, compact, and accurate appreciation of the information contained in an arbitrarily chosen document? We believe that the answer to this question is “yes”, and in the following sections, we present an overview of how term identification can be augmented and extended for the purpose of content characterisation.

3 Extended phrasal analysis and anaphora resolution

The questions raised at the end of the previous section concern the wider applicability of linguistic processing targeted at term identification. Three problems arise when “vanilla” term sets are considered as the basis for a content characterisation task.

3.1 Terms as content indicators

The first is, broadly construed, a problem of undergeneration. For a set of phrases to be truly representative of document content, it must provide an exhaustive description of the entities discussed in the text. That is, it must contain not just those expressions which satisfy the strict phrasal definition of “technical term”, but rather every expression which mentions a participant in the events described in the text. Such broad coverage is precisely *not* the goal of canonical term identification, which extracts only those expressions that have a suitably rich amount of descriptive content (compound nominals and nominals plus modifiers), ignoring e.g. pronouns and reduced descriptions. Phrasal analysis must therefore be extended to include (at least) all the nominal expressions in a text.

Extending phrasal analysis in this way, however, exacerbates a problem already noted: a full listing of all the terms that occur in a text, even when attention is restricted to technical terms in the strict sense, is typically too large to be usefully presented as a representation of a document’s content. Thus the second problem when using term sets as a basis for content characterisation is one of overgeneration: presentation of a list of phrases whose size rapidly leads to information overload. A system that extracts phrases on the basis of relaxed canonical terminology constraints, without recourse to domain or genre restrictions that might help to limit the size of the term set (a constraint imposed by the goal of constructing a system that works on arbitrary texts), will typically generate a term set far larger than a user can absorb. What is needed, then, is some means of establishing referential links be-

tween phrases, thereby reducing a large phrase set to just those that *uniquely* identify the participants in the discourse.

The final problem is one of differentiation. While lists of terms such as the ones presented above (Justeson & Katz 1995, appendix) might be topical for the particular source document in which they occur, other documents within the same domain are likely to yield similar, overlapping sets of terms. (This is precisely the reason why technical term sets are not necessarily readily usable for document retrieval.) The result is that two documents containing the same or similar terms could be incorrectly classified as “about the same thing”, when in fact they focus on completely different subtopics within a shared domain. In order to resolve this problem, it is necessary to differentiate term sets not only according to their membership, but also according to the relative representativeness (of document content) of the terms they contain.

Although we approach these three problems in different ways, the solutions are related, and it is this inter-relation that provide the basis for constructing capsule overviews from phrasal analysis. The mechanisms involved in the construction of capsule overviews from a term set—in effect, the solution to the problem of differentiation—are described in Section 4. In the remainder of this section, we focus on the modifications and extensions to traditional term identification technology that are needed in order to use term sets as sources for content characterisation in the first place. These modifications solve the first of the two problems listed above.

3.2 Term sets and coreference classes

The problem of undergeneration is resolved by implementing a suitable generalisation—and relaxation—of the notion of a term, so that identification and extraction of phrasal units involves a procedure essentially like TERMS (Justeson & Katz 1995), but results in an exhaustive listing of all of the nominal expressions in the text. This is accomplished by running a phrasal grammar over text that has been analyzed by the LINGSOFT supertagger (Karlsson *et al.*, 1995), which provides information about the part of speech, number, gender, and grammatical function (as well as other features) of tokens in a text. The phrasal grammar targets expressions that consist of a head noun preceded by some number (possibly zero) of pre-nominal modifiers (nouns or adjectives). As a result, it extracts not just the complex nominals that meet the formal definition of technical terms, but reduced descriptions and pronouns as well. Phrasal analysis yields the set of all nominal expressions occurring in the text, which we refer to as an *extended phrase set*.

In order to eliminate the problem of overgeneration, it is necessary to reduce the extended phrase set to a smaller set of expressions which uniquely identify the objects referred to in the text, hereafter a *referent set*. We make the simplifying assumption that every

phrase identified by extended phrasal analysis constitutes a “mention” of a participant in the discourse (see Mani & MacMillan, 1996, for discussion of the notion of mention in the context of proper name interpretation); in order to construct a referent set, it is necessary to determine which expressions constitute mentions of the same referent.

Coreference is established largely through the application of an anaphora resolution procedure that is based on the algorithm developed by Lappin & Leass (1994). The fundamental difference between our algorithm (described in detail in Kennedy & Boguraev 1996a, 1996b) and the one developed by Lappin and Leass is that it is designed to provide a reliable interpretation from a considerably shallower linguistic analysis of the input.³ This constraint is imposed by the type of analysis we are working with—the shallow analysis provided by LINGSOFT and the set of terms constructed from it, structured only according to precedence relations, not hierarchical relations—which is dictated by our goal of extending content characterisation to arbitrary types of text documents.

The basic approach to anaphora resolution, however, is the same. The interpretation procedure involves moving through the text sentence by sentence and analysing the nominal expressions in each sentence from left to right (expressions identified by the phrasal grammar are marked both for overall position in the text and for the sentence in which they occur). There are two possible outcomes of this examination. Either an expression is identified as a mention of a new participant in the discourse, or it is taken to refer to a previously mentioned referent—i.e., it either introduces a new referent or is identified as coreferential with some other expression in the text.

Coreference is determined by a three step procedure. First, a set of candidate antecedents is collected, which includes all nominals within a local segment of discourse. Second, those expressions with which an anaphoric expression cannot possibly corefer, by virtue of morphological mismatch or syntactic restrictions, are eliminated from consideration.⁴ Finally, the remaining candidates are ranked according to their relative *salience* in the discourse (see below), and the most salient candidate is selected as the antecedent for the anaphor. (In the event that a coreference link cannot be established to some other expression, the nominal is taken to introduce a new referent.) Linguistic expressions that are identified as coreferential are grouped into equivalence classes, or *coreference classes*, and each coreference class is taken to represent a unique referent

³The Lappin-Leass algorithm works from the analysis provided by the McCord Slot Grammar parser (McCord 1990); our algorithm achieves comparable results on the basis of the LINGSOFT analysis (Karlsson *et al.* 1995); see (Kennedy & Boguraev 1996a) for a comparison.

⁴For discussion of how syntactic relations are inferred on the basis of a shallow linguistic analysis, see (Kennedy & Boguraev 1996a).

in the discourse. For any text, the set of such coreference classes constitutes its reference set.

A crucial component of this anaphora resolution procedure is the computation of a salience measure for terms that are identified as candidate antecedents for an anaphoric expression. This measure, which we refer to as *local salience*, is straightforwardly determined as a function of how a candidate satisfies a set of grammatical, syntactic, and contextual parameters, or “salience factors” (this term is borrowed from Lappin & Leass 1994). Individual salience factors are associated with numerical values, as shown below.⁵

SENT(<i>term</i>)	= 100 iff <i>term</i> is in the current sentence
CNTX(<i>term</i>)	= 50 iff <i>term</i> is in the current discourse segment
SUBJ(<i>term</i>)	= 80 iff <i>term</i> is a subject
EXST(<i>term</i>)	= 70 iff <i>term</i> is in an existential construction
POSS(<i>term</i>)	= 65 iff <i>term</i> is a possessive
ACC(<i>term</i>)	= 50 iff <i>term</i> is a direct object
DAT(<i>term</i>)	= 40 iff <i>term</i> is an indirect object
OBLQ(<i>term</i>)	= 30 iff <i>term</i> is the complement of a preposition
HEAD(<i>term</i>)	= 80 iff <i>term</i> is not contained in another phrase
ARG(<i>term</i>)	= 50 iff <i>term</i> is not contained in an adjunct

The local salience of a candidate is the sum of the values of the salience v factors that are satisfied by some member of the coreference class to which the candidate belongs; values may be satisfied at most once by each member of the class.

The most important aspect of this characterisation of local salience for our purposes is that the numerical values associated with the salience factors correspond to a relational structure that is directly computable on the basis of grammatical information about particular terms. This relational structure in turn provides the basis for ordering candidate antecedents according to their relative salience in some local segment of discourse, and (by hypothesis) their likelihood as antecedents for a pronoun.⁶

The overall success of anaphora resolution procedures built on top of such a measure (Lappin & Leass 1994 report 85% accuracy; our system, built on top of a shallower linguistic analysis, runs at 75% accuracy; see Kennedy & Boguraev, 1996a) provides evidence of its usefulness in the domain of anaphora resolution.

A much broader consequence of this approach to anaphora resolution, and one of particular relevance to the task at hand, is that it introduces both a working definition of salience and a mechanism for determining the salience of particular linguistic expressions

⁵Our salience factors mirror those used by Lappin and Leass, with the exception of POSS, which is sensitive to possessive expressions, and CNTX, which is sensitive to the discourse segment in which a candidate appears (see Section 4 below).

⁶The relational structure imposed by the values of the salience factors listed here is justified both linguistically, as a reflection of the functional hierarchy (Keenan & Comrie 1977), as well as by experimental results (Lappin & Leass 1994).

based on straightforwardly computable grammatical properties of terms. In the next section, we show how this measure can be extended for the purpose of salience-based content characterisation.

4 Salience-based content characterisation

Anaphora resolution solves a number of the problems that arise when term identification technology is extended to work on arbitrary texts. First, it reduces the total list of terms identified by extended phrasal analysis to just those that uniquely identify objects in the discourse. Second, it establishes crucial connections between text expressions that refer to the same entities. This latter result is particularly important, as it provides a means of “tracking” occurrences of prominent expressions throughout the discourse; see (Kennedy & Boguraev 1996b) for discussion of this point.

4.1 Topic stamps

The data reduction arising from distilling the extended phrase set down to a smaller referent set is still not enough, however. In order to further reduce the referent set to a compact, coherent, and easily absorbed listing of just those expressions which identify the most important objects in the text (i.e., in order to solve the third problem discussed above, that of differentiation), some additional structure must be imposed upon its members. One way to accomplish this is to rank the members of a referent set according to the relative prominence or importance in the discourse of the entities to which they refer—in other words, to order a term set according to the *salience* of its members.

Salience is a measure of the relative prominence of objects in discourse: objects at the centre of discussion have a high degree of salience; objects at the periphery have a correspondingly lower degree of salience. The hypothesis underlying salience-based content characterisation is that even though two related documents may instantiate the same term sets, if the documents are about different things, then the relative salience of the terms in the two documents should differ. If the relative salience of the members of a referent set can be determined, then, an ordering can be imposed which, in connection with an appropriate choice of threshold value, permits the reduction of the entire referent set to only those expressions that identify the most prominent participants in the discourse.

The reduced set of terms, in combination with information about local context at various levels of granularity (verb phrase, minimal clause, sentence, etc.) may then be folded into an appropriate presentation metaphor and displayed as a characterisation of a document’s content. Crucially, this type of analysis satisfies the important requirements of usability mentioned in Section 1.3: it is concise, it is coherent, and it does not introduce the cognitive overload associated with a full-scale term set. In a more general sense, this strategy for scaling up the phrasal analysis provided by

standard term identification technology has at its core the utilisation of a crucial feature of discourse structure: the prominence, over some segment of text, of particular referents—something that is missing from the traditional technology for ‘bare’ terminology identification.

Clearly, what is necessary to implement this type of approach to content characterisation is a means of computing the relative salience of the participants in a discourse as a function of the terms that refer to them. The hypothesis underlying the anaphora resolution procedure discussed in the previous section is that a measure of local salience, which reflects the prominence of an expression in some local segment of discourse, can be determined on the basis of the frequency of use and grammatical distribution of expressions in a text. Our proposal is that the same principles can be applied to determine a more global level of salience, which reflects the prominence of expressions across the entire discourse.

An important feature of local salience is that it is variable: the salience of a referent decreases and increases according to the frequency with which it is mentioned (by subsequent anaphoric expressions). When an anaphoric link is established, the anaphor is added to the equivalence class to which its antecedent belongs, and the salience of the class is boosted accordingly. If a referent ceases to be mentioned in the text, however, its local salience is incrementally decreased. This approach works well for the purpose of anaphora resolution, because it provides a realistic representation of the antecedent space for an anaphor by ensuring that only those referents that have mentions within a local domain have increased prominence. However, the goal of salience-based content characterisation differs from that of anaphora resolution in an important respect. In order to determine which linguistic expressions should be presented as broadly representative of the content of a document, it is necessary to generate a picture of the prominence of referents across the entire discourse, not just within a local domain.

For illustration of the intuition underlying this idea, consider the news article discussed in Section 1.3. Intuitively, the reason why ‘*priest*’ is the primary element of the title is that there are no less than eight references to the same actor in the body of the story (marked by italics in the example); moreover, these references occur in prominent syntactic positions: five are subjects of main clauses, two are subjects of embedded clauses, and one is a possessive. Similarly, the reason why ‘*Pope attack*’ is the secondary object of the title is that a constituent of the compound, ‘*Pope*’, also receives multiple mentions (five), although these references tend to occur in less prominent positions (two are direct objects).

In order to generate the broader picture of discourse structure needed to inform the selection of certain expressions as most salient, and therefore most representative of content, we introduce an elaboration of the local salience computation described above that

uses the same conditions to calculate a non-decreasing, global salience value for every referent in the text. This non-decreasing salience measure, which we refer to as *discourse salience*, reflects the distributional properties of a referent as the text story unfolds. In conjunction with the “tracking” of referents made available by anaphora resolution, discourse salience provides the basis for a coherent representation of discourse structure that indicates the topical prominence of specific terms in isolated segments of text. Most importantly, discourse salience provides exactly the information that is needed to impose the type of importance-based ranking of referents discussed above, which in turn provides the basis for the construction of capsule overviews out of referent sets (which are derived from terms sets, as discussed in Section 3.2). Specifically, by associating every referent with a discourse salience value, we can identify the topic stamps for a segment of text S as the n highest ranked objects in S , where n is a scalable value.

4.2 Discourse segments

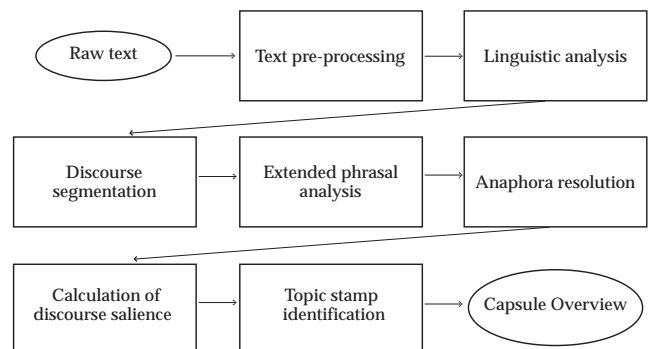
The notion “segment of text” plays an extremely important role in the content characterisation task, as it provides the basic units around which a capsule overview for a document is constructed. Again, the example from Section 1.3 provides a useful illustration of the important issues. The reason that the title of this passage works as an overview of its content is because the text itself is fairly short. As a text increases in length, the “completeness” of a short description as a characterisation of content deteriorates. If the intention is to use concise descriptions consisting of one or two salient phrases—i.e., topic stamps—along with information about the local context in which they appear as the primary information-bearing units for a capsule overview, then it follows that texts longer than a few paragraphs must be broken down into smaller units or “segments”.

In order to solve this problem, we recast a document as a set of *discourse segments*, which correspond to topically coherent, contiguous sections of text. The approach to segmentation we adopt implements a similarity-based algorithm along the lines of the one developed by (Hearst 1994), which identifies discourse segments text using a lexical similarity measure. By calculating the discourse salience of referents with respect to the results of discourse segmentation, each segment can be associated with a listing of those expressions that are most salient within the segment, i.e., each segment can be assigned a set of topic stamps. The result of these calculations, the set of segment-topic stamp pairs, ordered according to linear sequencing of the segments in the text, is the data structure on the basis of which the capsule overview for the entire document is constructed. In this way, the problem of content characterisation of a large text is reduced to the problem of finding topic stamps for each discourse segment.

4.3 Capsule overviews

To summarize, the approach to content characterisation that we have outlined here involves defining a suitable selection procedure, operating over a larger set of phrasal units than that generated by a typical term identification algorithm (including not only all terms, but term-like phrases, as well as their variants, reduced forms, and anaphoric references), with the following properties. First, it reduces this set to a list of expressions that uniquely refer to objects in the discourse (the referent set). Second, it makes informed choices about the degree to which each phrase is representative of the text as a whole. Finally, it presents its output in a form which retains contextual information for each phrase. The key to normalising the content of a document to a small set of distinguished, and discriminating, phrasal units is being able to establish a containment hierarchy of phrases (term-relational context-clause-sentence-paragraph-and so forth; this would eventually be exploited for capsule overview presentation at different levels of granularity), and being able to make refined judgements concerning the degree of importance of each unit, within some segment of text.

In simple terms, the goal is to filter a term set in such a way that those expressions which are identified as most salient are presented as representative of document content. This process of “salience-based content characterisation” builds on and extends the notion of salience that forms a crucial component of the anaphora resolution procedure developed by (Lappin & Leass 1994). Moreover, it presupposes very little in the way of linguistic processing, working solely on the basis of the shallow analysis provided by the LINGSOFT tagger. It thus meets the desired requirement of domain independence, permitting extension of the technology to a wide range of texts, without regard to genre, style, or source. The following diagram provides a schematic illustration of the primary components of the content characterisation procedure.



5 Example

We illustrate the procedure by highlighting certain aspects of a capsule overview of a recent *Forbes* article

(Hutheesing 1996). The document is of medium-to-large size (approximately four pages in print), and focuses on the strategy of Gilbert Amelio (former CEO of Apple Computer) concerning a new operating system for the Macintosh. Too long to quote here in full, the following passage from the beginning of the article contains the first, second and third segments, as identified by the discourse segmentation component described in Section 4.2, cf. (Hearst 1994); in the example below, segment boundaries are marked by extra vertical space).

"ONE DAY, everything Bill Gates has sold you up to now, whether it's Windows 95 or Windows 97, will become obsolete," declares Gilbert Amelio, the boss at Apple Computer. "Gates is vulnerable at that point. And we want to make sure we're ready to come forward with a superior answer."

Bill Gates vulnerable? Apple would swoop in and take Microsoft's customers? Ridiculous! Impossible! In the last fiscal year, Apple lost \$816 million; Microsoft made \$2.2 billion. Microsoft has a market value thirty times that of Apple.

Outlandish and grandiose as Amelio's idea sounds, it makes sense for Apple to think in such big, bold terms. Apple is in a position where standing pat almost certainly means slow death.

It's a bit like a patient with a probably terminal disease deciding to take a chance on an untested but promising new drug. A bold strategy is the least risky strategy. As things stand, customers and outside software developers alike are deserting the company. Apple needs something dramatic to persuade them to stay aboard. A radical redesign of the desktop computer might do the trick. If they think the redesign has merit, they may feel compelled to get on the bandwagon lest it leave them behind.

Lots of "ifs," but you can't accuse Amelio of lacking vision. Today's desktop machines, he says, are ill-equipped to handle the coming power of the Internet. Tomorrow's machines must accommodate rivers of data, multimedia and multitasking (juggling several tasks simultaneously).

We're past the point of upgrading, he says. Time to scrap your operating system and start over. The operating system is the software that controls how your computer's parts (memory, disk drives, screen) interact with applications like games and Web browsers. Once you've done that, buy new applications to go with the reengineered operating system.

Amelio, 53, brings a lot of credibility to this task. His resume includes both a rescue of National Semiconductor from near-bankruptcy and 16 patents, including one for co-inventing the charge-coupled device.

But where is Amelio going to get this new operating system? From Be, Inc., in Menlo Park, Calif., a half-hour's drive from Apple's Cupertino headquarters, a hot little company founded by ex-Apple visionary Jean-Louis Gasse. Its BeOS, now undergoing clinical trials, is that radical redesign in operating systems that Amelio is talking about. Married to hardware from Apple and Apple cloners, the BeOS just might be a credible competitor to Microsoft's Windows, which runs on IBM-compatible hardware.

The capsule overview was automatically generated by a fully implemented, and operational, system, which incorporates all of the processing components identified above. The relevant sections of the overview (for the three segments of the passage quoted) are listed below. We ignore here the issue of the right presentation metaphor for topic stamps (but see Boguraev *et al.*, 1998). The listing of topic stamps in context shown below provides the core data out of which a capsule overview is constructed; such a listing is not the most effective and informative presentation of the data, and should be regarded as indicative of the data structure underlying the overview.

1 APPLE; MICROSOFT

APPLE would swoop in and take MICROSOFT's customers?
 APPLE lost \$816 million;
 MICROSOFT made \$2.2 billion.

MICROSOFT has a market value thirty times that of APPLE
 it makes sense for APPLE
 APPLE is in a position
 APPLE needs something dramatic

2 DESKTOP MACHINES; OPERATING SYSTEM

Today's DESKTOP MACHINES, he [Gilbert Amelio] says
 Tomorrow's MACHINES must accommodate rivers of data
 Time to scrap your OPERATING SYSTEM and start over
 The OPERATING SYSTEM is the software that controls
 to go with the REENGINEERED OPERATING SYSTEM

3 GILBERT AMELIO; NEW OPERATING SYSTEM

AMELIO, 53, brings a lot of credibility to this task
 HIS [Gilbert Amelio] resumé includes
 where is AMELIO going to get this NEW OPERATING SYSTEM?
 radical redesign in OPERATING SYSTEMS that AMELIO is talking about

The division of this passage into segments, and the segment-based assignment of topic stamps, exemplifies a capsule overview's "tracking" of the underlying coherence of a story. The discourse segmentation component recognizes shifts in topic—in this example, the shift from discussing the relation between Apple and Microsoft to some remarks on the future of desktop computing to a summary of Amelio's background and plans for Apple's operating system. Layered on top of segmentation are the topic stamps themselves, in their relational contexts, at a phrasal level of granularity.

The first segment sets up the discussion by positioning Apple opposite Microsoft in the marketplace and focusing on their major products, the operating systems. The topic stamps identified for this segment, APPLE and MICROSOFT, together with their local contexts, are both indicative of the introductory character of the opening paragraphs and highly representative of the gist of the first segment. Note that the apparent unformativeness of some relational contexts, for example, '... APPLE is in a position ...', does not pose a serious problem. An adjustment of the granularity—at capsule overview presentation time—reveals the larger context in which the topic stamp occurs (e.g., a sentence), which in turn inherits the high topicality ranking of its anchor: 'APPLE is in a position where standing pat almost certainly means slow death.'

For the second segment of the sample, OPERATING SYSTEM and DESKTOP MACHINES have been identified as representative. The set of topic stamps and contexts illustrated provides an encapsulated snapshot of the segment, which introduces Amelio's views on coming challenges for desktop machines and the general concept of an operating system. Again, even if some of these are somewhat under-specified, more detail is easily available by a change in granularity, which reveals the definitional nature of the even larger context 'The OPERATING SYSTEM is the software that controls how your computer's parts...'

The third segment of the passage exemplified above is associated with the stamps GILBERT AMELIO and NEW OPERATING SYSTEM. The reasons, and linguistic rationale, for the selection of these particular noun

phrases as topical are essentially identical to the intuition behind ‘*priest*’ and ‘*Pope attack*’ being the central topics of the example in Section 1.3. The computational justification for the choices lies in the extremely high values of salience, resulting from taking into account a number of factors: co-referentiality between ‘*Amelio*’ and ‘*Gilbert Amelio*’, co-referentiality between ‘*Amelio*’ and ‘*His*’, syntactic prominence of ‘*Amelio*’ (as a subject) promoting topical status higher than for instance ‘*Apple*’ (which appears in adjunct positions), high overall frequency (four, counting the anaphor, as opposed to three for ‘*Apple*’—even if the two get the same number of text occurrences in the segment)—and boost in global salience measures, due to “priming” effects of both referents for ‘*Gilbert Amelio*’ and ‘*operating system*’ in the prior discourse of the two preceding segments. Even if we are unable to generate a single phrase summary in the form of, say, ‘*Amelio seeks a new operating system*’, the overview for the closing segment comes close; arguably, it is even better than any single phrase summary.

As the discussion of this example illustrates, a capsule overview is derived by a process which facilitates partial understanding of the text by the user. The final set of topic stamps is designed to be representative of the core of the document content. It is *compact*, as it is a significantly cut-down version of the full list of identified terms. It is highly *informative*, as the terms included in it are the most prominent ones in the document. It is *representative* of the whole document, as a separate topic tracking module effectively maintains a record of where and how referents occur in the entire span of the text. As the topics are, by definition, the primary content-bearing entities in a document, they offer *accurate* approximation of what that document is about.

6 Related and future work

Our framework clearly attempts to balance the conflicting requirements of the two primary approaches to the document summarisation task. By design, we target any text type, document genre, and domain of discourse, and thus compromise by forgoing in-depth analysis of the full meaning of the document. On the other hand, because our content characterisation procedure presents extracted elements in the contexts in which they appear in a text, it provides a less ambiguous abstraction of the core meaning of the text than the traditional passage extraction algorithms, which offer certain sentence- or paragraph-sized passages deemed indicative of content by means of similarity scoring metrics.

By choosing a phrasal—rather than sentence- or paragraph-based—granularity of representation, we can obtain a more refined view into highly relevant fragments of the source; this also offers a finer-grained control for adjusting the level of detail in capsule overviews. Exploiting a notion of discourse contiguity

and coherence for the purposes of full source coverage and continuous context maintenance ensures that the entire text of the document is uniformly represented in the overview. Finally, by utilising a strong linguistic notion of salience, the procedure can build a richer representation of the discourse objects, and exploit this for informed decisions about their prominence, importance, and ultimately topicality. These are notable characteristics of our notion of capsule overviews—to the extent that the focused and rich semantic information they encapsulate is sufficiently indicative of content to offset the (initial) unfamiliarity in dealing with phrasal-, rather than sentence-based, fragments.

At present, salience calculations are driven from contextual analysis and syntactic considerations focusing on discourse objects and their behaviour in the text.⁷ Given the power of our phrasal grammars, however, it is conceivable to extend the framework to identify, explicitly represent, and similarly rank, higher order expressions (e.g. events, or properties of objects). This may not ultimately change the appearance of a capsule overview; however, it will allow for even more informed judgements about relevance of discourse entities. More importantly, it is a necessary step towards developing more sophisticated discourse processing techniques (such as those discussed in Sparck Jones, 1993b), which are ultimately essential for the automatic construction of true summaries.

Currently, we analyse individual documents; unlike (McKeown & Radev 1995), there is no notion of calculating salience across the boundaries of more than one document—even if we were to know in advance that they are somehow related. However, we are experimenting with using topic stamps as representation and navigation “labels” in a multi-document space; we thus plan to fold in awareness of document boundaries (as an extension to tracking the effects of discourse segment boundaries within a single document).

The approach presented here can be construed, in some sense, as a type of passage extraction, it is however considerably less exposed to problems like pronouns out of context, or discontinuous sentences presented as contiguous passages; cf. (Paice 1990). This is a direct consequence of the fact that we employ anaphora resolution to construct a discourse model with explicit representation of objects, and use syntactic criteria to extract coherent phrasal units. For the same reason, topic stamps are quantifiably adequate content abstractions: see (Kennedy & Boguraev 1996a), and Section 3.2 above, for evaluation of the anaphora resolution algorithm. Following deployment of capsule overviews in an environment designed to facilitate rapid on-line skimming of news stories (Houde, Bellamy, & Leahy 1998), we are also in the process of

⁷Issues arising from differences in the semantic types of phrases—e.g. definite vs. indefinite descriptions—are currently ignored; the extent to which these might impact a procedure like ours is an open question.

designing a user study to determine the utility, from usability point of view, of capsule overviews as defined here.

Recent work in summarisation has begun to focus more closely on the utility of document fragments with granularity below that of a sentence. For example, (McKeown & Radev 1995) pro-actively seek, and use to great leverage, certain cue phrases which denote specific rhetorical and/or inter-document relationships. (Mahesh 1997) uses phrases as "sentence surrogates", in a process called sentence simplification; his rationale is that with hypertext, a phrase can be used as a place-holder for the complete sentence, and/or is a more conveniently manipulated, compared to a sentence. Even in passage extraction work, notions of multi-word expressions have found use as one of several features driving a statistical classifier scoring sentences for inclusion in a sentence-based summary (Kupiec, Pedersen, & Chen 1995). In all of these examples, the use of a phrase is somewhat peripheral to the fundamental assumptions of the particular approach; more to the point, it is a different kind of object that the summary is composed from (a template, in the case of McKeown & Radev 1995), or that the underlying machinery is seeking to identify (sentences, in the case of Mahesh 1997, and Kupiec, Pedersen, & Chen 1995). In contrast, our adoption of phrasal expressions as the atomic building blocks for capsule overviews is central to the design; it drives the entire analysis process, and is the underpinning for our discourse representation.

References

- AAAI 1998 Spring Symposium Series. 1998. *Intelligent Text Summarization (Working Papers)*, Stanford, California.
- Advanced Research Projects Agency. 1993a. *Fifth Message Understanding Conference (MUC-5)*, Baltimore, Maryland: Software and Intelligent Systems Technology Office.
- Advanced Research Projects Agency. 1993b. *Tipster Text Program: Phase I*, Fredericksburg, Virginia.
- Barzilay, R., and Elhadad, M. 1997. Using lexical chains for text summarization. In *Proceedings of ACL'97 Workshop on Intelligent, Scalable Text Summarization*, 10-17.
- Boguraev, B.; Wong, Y. Y.; Kennedy, C.; Bellamy, R.; Brawer, S.; and Swartz, J. 1998. Dynamic presentation of document content for rapid on-line browsing. In *AAAI Spring Symposium on Intelligent Text Summarization (Working Papers)*, 118-128.
- Bourigault, D. 1992. Surface grammatical analysis for the extraction of terminological noun phrases. In *15th International Conference on Computational Linguistics*.
- Brandow, R.; Mitze, K.; and Rau, L. 1995. Automatic condensation of electronic publications by sentence selection. *Information Processing and Management* 31(5):675-685.
- Caruso, D. 1997. New software summarizes documents. *The New York Times*.
- Dagan, I., and Church, K. 1995. Termight: identifying and translating technical terminology. In *4th Conference on Applied Natural Language Processing*.
- Defense Advanced Research Projects Agency. 1992. *Fourth Message Understanding Conference (MUC-4)*, McLean, Virginia: Software and Intelligent Systems Technology Office.
- Defense Advanced Research Project Agency. 1998. TIPSTER/SUMMAC *Summarization Analysis; Tipster Phase III 18-Month Meeting*, NIST, Fairfax, Virginia.
- DeJong, G. 1982. An overview of the FRUMP system. In Lehnert, W., and Ringle, M., eds., *Strategies for Natural Language Parsing*. Hillsdale, NJ: Lawrence Erlbaum Associates. 149-176.
- Endres-Niggemeyer, B. 1998. A grounded theory approach to expert summarizing. In *Proceedings of AAAI Spring Symposium on Intelligent Text Summarization (Working Papers)*, 140-142.
- Hearst, M. 1994. Multi-paragraph segmentation of expository text. In *32nd Annual Meeting of the Association for Computational Linguistics*.
- Houde, S.; Bellamy, R.; and Leahy, L. 1998. In search of design principles for tools and practices to support communication within a learning community. *SIGCHI Bulletin* 30(2).
- Hovy, E., and Lin, C. Y. 1997. Automated text summarization in SUMMARIST. In *Proceedings of ACL'97 Workshop on Intelligent, Scalable Text Summarisation*, 18-24.
- Hutheesing, N. 1996. Gilbert Amelio's grand scheme to rescue Apple. *Forbes Magazine*.
- Justeson, J. S., and Katz, S. M. 1995. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering* 1(1):9-27.
- Karlsson, F.; Voutilainen, A.; Heikkilä, J.; and Antilla, A. 1995. *Constraint grammar: A language-independent system for parsing free text*. Berlin/New York: Mouton de Gruyter.
- Keenan, E., and Comrie, B. 1977. Noun phrase accessibility and universal grammar. *Linguistic Inquiry* 8:62-100.
- Kennedy, C., and Boguraev, B. 1996a. Anaphora for everyone: Pronominal anaphora resolution without a parser. In *Proceedings of COLING-96 (16th International Conference on Computational Linguistics)*.
- Kennedy, C., and Boguraev, B. 1996b. Anaphora in a wider context: Tracking discourse referents. In Wahlster, W., ed., *Proceedings of ECAI-96 (12th European Conference on Artificial Intelligence)*. Budapest, Hungary: John Wiley and Sons, Ltd, London/New York.
- Kupiec, J.; Pedersen, J.; and Chen, F. 1995. A trainable document summarizer. In *Proceedings of the 18th An-*

- nual International ACM SIGIR Conference on Research and Development in Information Retrieval, 68–73.
- Lappin, S., and Leass, H. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics* 20(4):535–561.
- Luhn, H. 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development* 2:159–165.
- Mahesh, K. 1997. Hypertext summary extraction for fast document browsing. In *Proceedings of AAAI Spring Symposium on Natural Language Processing for the World Wide Web*, 95–104.
- Mani, I., and MacMillan, T. 1996. Identifying unknown proper names in newswire text. In Boguraev, B., and Pustejovsky, J., eds., *Corpus Processing for Lexical Acquisition*. Cambridge, Mass: MIT Press. 41–59.
- Marcu, D. 1997. From discourse structures to text summaries. In *Proceedings of ACL'97 Workshop on Intelligent, Scalable Text Summarisation*, 82–88.
- McCord, M. M. 1990. Slot grammar: a system for simpler construction of practical natural language grammars. In Studer, R., ed., *Natural language and logic: international scientific symposium*, Lecture Notes in Computer Science. Berlin: Springer Verlag. 118–145.
- McKeown, K., and Radev, D. 1995. Generating summaries of multiple news articles. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 74–82.
- Paice, C. D. 1990. Constructing literature abstracts by computer: techniques and prospects. *Information Processing and Management* 26:171–186.
- Preston, K., and Williams, S. 1994. Managing the information overload: new automatic summarization tools are good news for the hard-pressed executive. *Physics in Business*.
- Rau, L. 1988. Conceptual information extraction and retrieval from natural language input. In *Proceedings of RIAO-88, Conference on User-oriented Content-Based Text and Image Handling*, 424–437.
- Reimer, U., and Hahn, U. 1997. A formal model of text summarization based on condensation operators of a terminological logic. In *Proceedings of ACL'97 Workshop on Intelligent, Scalable Text Summarisation*.
- Resnik, P. 1997. Evaluating multilingual gisting of web pages. In *AAAI Spring Symposium on Natural Language Processing and the World-Wide Web (Working Papers)*.
- Salton, G.; Singhal, A.; Buckley, C.; and Mitra, M. 1996. Automatic text decomposition using text segments and text themes. In *Seventh ACM Conference on Hypertext*.
- Sparck Jones, K. 1993a. Discourse modelling for automatic text summarising. Technical Report 290, University of Cambridge Computer Laboratory, Cambridge, England.
- Sparck Jones, K. 1993b. What might be in a summary? In Knorz; Krause; and Womser-Hacker., eds., *Information Retrieval 93: Von der Modellierung zur Anwendung*, 9–26.
- Sparck Jones, K. 1997. Summarising: Where are we now? Where should we go? In *Keynote address to ACL'97 Workshop on Intelligent, Scalable Text Summarisation*.
- Tait, J. 1983. *Automatic summarising of English texts*. Ph.D. Dissertation, University of Cambridge Computer Laboratory, Cambridge, England. Technical Report 47.