Salience-based Content Characterisation of Text Documents

Branimir Boguraev
Apple Research Laboratories
Apple Computer, Inc.
bkb@research.apple.com

Christopher Kennedy
Department of Linguistics
Northwestern University
kennedy@babel.ling.nwu.edu

23 July, 1997

Abstract

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 salience, as a more appropriate and representative measure of a document's "aboutness".

Salience-based Content Characterisation of Text Documents

Abstract

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 salience, as a more appropriate and representative measure of a document's "aboutness".

1 Introduction

1.1 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 [7] and Tait [29]; more recently, the DARPA-sponsored TIPSTER programme ([2])—and, in particular, the message understanding conferences (MUC: e.g. [6] and [1])—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. [21]), 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 Spärck Jones, [27] and [28] 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, [17]), 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 and Williams [23], Rau *et al.* [25]); how to derive more representative scoring functions, e.g. for complex documents, such as multi-topic ones (Salton *et al.* [26]), or where training from professionally prepared abstracts is possible (Kupiec *et al.* [15]); what heuristics might be developed for improving readability and coherence of "narratives" made up of discontiguous source document chunks (Paice [22]); or with optimal presentations of such passage extracts, aimed at retaining some sense of larger and/or global context (Mahesh [18]).

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 (Rau [24]), especially with the recent wave of commercial products announcing built-in "summarisation" (by extraction) features (Caruso [4]).[1]

1.2 Capsule overviews

In this work, we take an approach which might be construed as striving for the best of both worlds. We use linguistically-intensive techniques to identify highly salient phrasal units across the entire span of the document, capable of functioning as representations of a document's content. The set of salient 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 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.[2] In particular, we seek to characterise a document's content in a way which is representative of the full flow of

[1] Also at `http://www.nytimes.com/library/cyber/digicom/012797digicom.html`

[2] A list of topic stamps is, by itself, not a coherent summary; however, in conjunction with an appropriately designed presentation metaphor—aiming, overall, to retain contextual cues associated with topic stamps as they appear in the text—our topic stamps provide a more informative representation of content than just a list of (noun or verb) phrases. This paper focuses on the linguistic processes underlying the automatic identification and extraction of topic stamps and their organisation within capsule overviews. 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.

the narrative; this is in contrast to passage extraction methods, which typically highlight only certain fragments (an unavoidable consequence of the compromises necessary when the passages are sentence-sized).

A capsule overview is not a fully instantiated meaning template, either. 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.

For us, a capsule overview is instead a coherently presented list of the linguistic expressions which refer 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:[3]

[3]Adapted from an example of S. Nirenburg.

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. Our strategy is to mine a document for the most salient—and by hypothesis, the most representative—phrasal units,

as well as the relational expressions they are associated with, with the goal of establishing the kind of core content specification that is captured by the title of this example. The goal of this paper is to describe a procedure that implements this selective mining of a document for 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 identification, we outline in section 2 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 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 3. Section 4 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 Phrasal identification

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, capable of term identification across a wide range of techical prose, that maintain their quality regardless of document domain and type. Since topic stamps are essentially phrasal units with certain

discourse properties—they manifest a high degree of salience within contiguous discourse segments—we define the task of content characterisation as one of identifying phrasal units with lexico-syntactic properties similar to those of technical terms and with discourse properties which signify their status as "most prominent". In section 3, we show how these discourse properties are computable as a function of the grammatical distribution of the phrase. Below we discuss the application of terminology identification to the content characterisation task.

2.1 Technical terminology: strengths and limitations

One of the best defined procedures for technical terminology identification is the TERMS developed by Justeson and Katz [10], 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 [10], part-of-speech tagging [5], or syntactic parsing [20]), and it has strong cross-linguistic application (see, for instance, [3]). 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 the terms which are representative of the "aboutness" of a document, as the expressions that provide important information about an individual document's content often do not distinguish that document from other texts within the same domain.

Still, is is clear that a program like TERMS is a good starting point for distilling representative lists. For example, Justeson and Katz [10, appendix] present 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.

Unfortunately, although such lists are representative, they 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 "term" sets derived from arbitrary texts can provide the same level of informativeness as term sets derived from technical documents, such as those listed above. Certainly, we cannot even talk of "technical terms" in the narrower sense assumed by the TERMS algorithm. This raises the following question: can similar phrase identification technology be used to construct phrase sets which can be construed as broadly characteristic of the topical content of a document? In other words, 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? Below we address this issue by showing how a basic term set can be enriched and structured in order to convey a more refined picture of content.

2.2 Extended phrasal analysis

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

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 should 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. Phrasal analysis must therefore be extended to include pronouns and reduced descriptions, in addition to the more complex nominals which satisfy the strict definition of technical term.

Overgeneration As noted above, a full listing of technical terms in even the strict sense is typically too large to be usefully presented as a representation of a document's content; relaxation of the canonical phrasal definition of technical term as described above only exacerbates this problem, rapidly leading to information overload. As a result, when applied to a document without regard to domain or genre, a system which extracts phrases on the basis of relaxed canonical terminology constraints will typically generate a term set far larger than a user can absorb. At the same time, the set may contain several distinct

phrasal units which refer to the same discourse object. In order to capture these crucial connections, some means of establishing anaphoric links is required.

Differentiation Finally, while a list of terms may 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. Unacceptably, this might result in two documents containing the same or similar terms being classified as "about the same thing", when in fact they might focus on completely different subtopics within a shared domain.

Although we approach these three problems in slightly different ways, the solutions are interconnected, and it is their interaction that is important to the construction of capsule overviews from a phrasal analysis. The exact mechanisms involved in the construction of a capsule overview from a term set are described in more detail in section 3; here we outline the modifications and extensions to traditional term identification technology which address the problems listed above.

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 [10], but results in an *extended phrase set*, which contains an exhaustive listing of the nominal expressions in the text.

The problem of overgeneration is resolved by reducing the extended phrase set to a smaller set of expressions which *uniquely* identify the objects referred to in the text (hereafter a *referent set*) through the application of an anaphora resolution procedure (this is discussed in detail in section 3.1 below). Anaphora resolution solves both of the problems noted above: it reduces the total list of terms (in the relaxed sense) to just those that uniquely identify objects in the discourse, and it establishes crucial connections between text expressions that refer to the same entities. This latter result is particularly important, as it pro-

vides a means of "tracking" occurrences of prominent expressions throughout the discourse (see Kennedy and Boguraev [14] for discussion of this point).

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 small, coherent, and easily absorbed listing of just those expressions which identify the most important objects in the text, its members must be ranked according to a measure of the importance, or *salience*, in the discourse of the entities to which they refer. As we will show in section 3.2, such a ranking provides the basis for the identification of topic stamps.

Ranking by salience also solves the third problem discussed above, that of differentation. Although 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 will differ as a function of differences in use and grammatical distribution. The underlying intuition is that term sets can be differentiated in two ways: lexically, by virtue of containing different terms, or structurally, by virtue of the ordering of their members. Ordered term sets provide distinct characterisations of documents, even if the overall lexical content of the terms is uniform. Given a formalised and computable notion of salience that accurately represents the relative prominence of the participants in a discourse, then, we can overcome the lack of coherence inherent in unstructured term sets by imposing an ordering on the terms according to the salience of the entities to which they refer.

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), that makes informed choices about the degree to which each phrase is representative of the text as a

whole, and presents its output in a form which retains contextual information for each phrase. On this view, 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 (which 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. The next section shows how this process of "salience-based content characterisation" can be implemented by building on and extending the notion of salience used that forms a crucial component of the anaphora resolution procedure developed by Lappin and Leass [16].

3 Salience-based content characterisation

Salience is a measure of the relative prominence of objects in discourse: objects with high salience are the focus of attention; those with low salience are at the periphery. As discussed in the previous section, in an effort to resolve the problems facing a term-based approach to content characterisation, we have developed a procedure which uses a salience feature as the basis for a "ranking by importance" of an unstructured referent set, and ultimately fortopic stamp identification. By determining the salience of the members of a referent set, 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. This 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 analysis satisfies the important requirements of usability mentioned above: it is consise, 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.

The following sections describe the main details of our technology. First, we show how the procedure for anaphora resolution that we adopt in order to reduce the extended phrase set to a referent set, as described above, provides a method for computing the salience based on linguistic information readily available to scalable and robust identification technologies. Second, we introduce an extended notion of salience that, when applied to arbitrary sets of phrasal units, generates an ordering that accurately represents the relative prominence of the objects referred to in a document. Finally, we explain more concretely what we mean by "segment of text", why segments are important, and how they are determined. In section 4, we turn to a discussion of an example that provides an overview of a linguistic processing environment that carries out these tasks, while remaining open-ended with respect to the language, domain, style and genre of the texts we want to be able to handle.

3.1 Anaphora resolution and local salience

As noted in section 2.2, the set of expressions generated by extended phrasal analysis typically contains a number of anaphoric expressions—pronouns, reduced descriptions, etc.—which must be associated with content-bearing antecedents. The anaphora resolution algorithm we employ to achieve this goal

is based on a procedure developed by Lappin and Leass [16], and is described in detail in Kennedy and Boguraev [13], [14]. The fundamental difference between our algorithm and the one described by Lappin and Leass is that it is designed to provide a reliable interpretation from a considerably shallower linguistic analysis of the input, a constraint which is imposed by our goal of achieving content characterisation for arbitrary types of text documents.[4] We make the simplifying assumption that every phrase identified by extended phrasal analysis constitutes a "mention" of a participant in the discourse (see Mani and Macmillan [19] for discussion of the notion of "mention" in the context of proper names interpretation), and anaphora resolution is geared towards determining which expressions constitute mentions of the same referent. To this end, linguistic expressions that are identified as coreferential are grouped into equivalence classes, and each equivalence class is taken to represent a unique referent in the discourse. The set of such equivalence classes constitutes the referent set discussed above.

The immediate result of anaphora resolution is to reduce the extended phrase set; the larger consequence is that it provides the basis for the identification of topic stamps, as it introduces both a working definition of salience and a formal mechanism for determining the salience of particular linguistic expressions. Roughly speaking, the anaphora resolution procedure locates an antecedent for an anaphoric expression by first eliminating all impossible candidate antecedents, then ranking the remaining candidates according to a salience measure and selecting the most salient candidate as the antecedent. This measure, which we refer to as *local salience*, is a function of how a candidate satisfies a set of grammatical, syntactic, and contextual parameters. Following Lappin and

[4]The Lappin-Leass algorithm works off of the analysis provided by the McCord Slot Grammar parser [20]; our algorithm achieves comparable results on the basis of the analysis provided by the LINGSOFT supertagger [11].

Leass, we refer to these constraints as "salience factors". Individual salience factors are associated with numerical values, as shown below.[5]

SENT: 100 iff the expression is in the current sentence.
CNTX: 50 iff the expression is in the current discourse segment.
SUBJ: 80 iff the expression is a subject.
EXST: 70 iff the expression is in an existential construction.
POSS: 65 iff the expression is a possessive.
ACC: 50 iff the expression is a direct object.
DAT: 40 iff the expression is an indirect object.
OBLQ: 30 iff the expression is the complement of a preposition.
HEAD: 80 iff the expression is not contained in another phrase.
ARG: 50 iff the expression is not contained in an adjunct.

The local salience of a candidate is the sum of the values of the salience factors that are satisfied by some member of the equivalence class to which the candidate belongs; values may be satisfied at most once by each member of the class. The most important aspect of these numerical values for our concerns is that they impose a relational structure on the salience factors, which in turn provides the basis for ordering referents according to their relative prominence in the disourse.[6]

3.2 Discourse salience and capsule overview

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

[5]Our salience factors mirror those used by Lappin and Leass, with the exception of POSS, which is senstive to possessive expressions, and CNTX, which is sensitive to the discourse segment in which a candidate appears (see section 3.2 below).

[6]The relational structure imposed by the values of the salience factors listed here ordering is justified both linguistically, as a reflection of the functional hierarchy (see e.g. Keenan and Comrie [12]), as well as by experimental results (see Lappin and Leass [16], Kennedy and Boguraev [13], [14] for discussion).

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. The ultimate goal of salience-based content characterisation differs from that of anaphora resolution in an important respect, however. 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 picure 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. Intuitively, the reason why *'priest'* is at the focus 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 expresssions 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 calcluate 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 distribu-

tional properties of a referent as the text story unfolds. In conjunction with the "tracking" of referents made avaliable by anaphora resolution, discourse salience provides the basis for a coherent representation of discourse structure that indicates the topical prominence of individual mentions of referents 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 in section 2.2 that is required for the identification of topic stamps. 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 referents in `S`, where `n` is a scalable value.

The notion "segment of text" plays an extremely important role in the content characterisation task, as it is the provides the basic units around which a capsule overview for a document is constructed. Again, the example from section 1 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 [8], 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, a the set of segment-topic stamp pairs, ordered according to linear sequencing of the segments in the text, can then be returned as the capsule overview for the entire document. 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 Example

The operational components of salience-based content characterisation fall in the following categories: discourse segmentation; phrasal analysis (of nominal expressions and relations); anaphora resolution and generation of the referent set; calculation of discourse salience and identification of topic stamps; and enriching topic stamps with information about relational context(s). Some of the functionality follows directly from terminology identification; in particular, both relation identification and extended phrasal analysis are carried out by running a phrasal grammar over a stream of text tokens tagged for morphological, syntactic, and grammatical function; this is in addition to a grammar mining for terms and, generally, referents. (Base level linguistic analysis is provided by the LINGSOFT supertagger, [11].) The later, more semantically-intensive algorithms are described in some detail in [13] and [14].

We illustrate the procedure by highlighting certain aspects of a capsule overview of a recent *Forbes* article ([9]). 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 3.2 (cf. [8]); 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 coinventing 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 Gassee. 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.[7]

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 accomodate 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

[7]We ignore here the issue of the right presentation metaphor for topic stamps. The listing of topic stamps in context shown here provides the core data out of which a capsule overview is constructed; such a listing is arguably not the most effective and informative presentation of the data, however.

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 uninformativeness 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 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. 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.

5 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, our content characterisation procedure remains closer to the core meaning than the approximations offered by traditional passage extraction algorithms, with certain sentence- or paragraph-sized passages deemed indicative of content by means of similarity scoring metrics.

By choosing a phrasal granularity of representation—rather than sentence- or paragraph-based—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; salience thus becomes central to deriving a strong sense of a document's "aboutness".

At present, salience calculations are driven from contextual analysis and syntactic considerations focusing on discourse objects and their behaviour in the text. 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 Spärck Jones [28]), which are ultimately essential for the automatic construction of true summaries.

Currently, we analyse individual documents; unlike McKeown and Radev [21], 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). Even though the approach presented here can be construed, in some sense, as a type of passage extraction, it is considerably less exposed to problems like pronouns out of context, or discontinuous sentences presented as contiguous passages (cf. Paice [22]). 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 and Boguraev [13] for evaluation of the anaphora resolution algorithm. We are also in the process of 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 closer on the utility of document fragments with granularity below that of a sentence. Thus McKeown and Radev [21] pro-actively seek, and use to great leverage, certain cue phrases which denote specific rhetorical and/or inter-document relationships. Mahesh [18] 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 classifer scoring sentences for inclusion in a sentence-based summary (Kupiec *et al.* [15]). 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 [21]), or that the underlying machinery is seeking to identify (sentences, in the case of [18] and [15]). 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

[1] Advanced Research Projects Agency. *Fourth Message Understanding Conference (*MUC-5*)*, Baltimore, Maryland, 1993. Software and Intelligent Systems Technology Office.

[2] Advanced Research Projects Agency. *Tipster Text Program: Phase I*, Fredericksburg, Virginia, 1993.

[3] D. Bourigault. Surface grammatical analysis for the extraction of terminological noun phrases. In *14th International Conference on Computational Linguistics*, Nantes, France, 1992.

[4] D. Caruso. New software summarizes documents. The New York Times, January 27, 1997.

[5] I. Dagan and K. Church. Termight: identifying and translating technical terminology. In *4th Conference on Applied Natural Language Processing*, Stuttgart, Germany, 1995.

[6] Defense Advanced Research Projects Agency. *Fourth Message Understanding Conference (*MUC-4*)*, McLean, Virginia, 1992. Software and Intelligent Systems Technology Office.

[7] G. DeJong. An overview of the FRUMP system. In W. Lehnert and M. Ringle, editors, *Strategies for Natural Language Parsing*, pages 149–176. Lawrence Erlbaum Associates, Hillsdale, NJ, 1982.

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

[9] N. Hutheesing. Gilbert Amelio's grand scheme to rescue Apple. Forbes Magazine, December 16, 1996.

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

[11] F. Karlsson, A. Voutilainen, J. Heikkilä, and A. Antilla. *Constraint grammar: A language-independent system for parsing free text*. Mouton de Gruyter, Berlin/New York, 1995.

[12] E. Keenan and B. Comrie. Noun phrase accessibility and universal grammar. *Linguistic Inquiry*, 8:62–100, 1977.

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

[14] C. Kennedy and B. Boguraev. Anaphora in a wider context: Tracking discourse referents. In W. Wahlster, editor, *Proceedings of ECAI-96 (12th European Conference on Artificial Intelligence)*, Budapest, Hungary, 1996. John Wiley and Sons, Ltd, London/New York.

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

[16] S. Lappin and H. Leass. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–561, 1994.

[17] H. Luhn. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2:159–165, 1959.

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

[19] I. Mani and T. MacMillan. Identifying unknown proper names in newswire text. In B. Boguraev and J. Pustejovsky, editors, *Corpus Processing for Lexical Acquisition*, pages 41–59. MIT Press, Cambridge, Mass, 1996.

[20] M. M. McCord. Slot grammar: a system for simpler construction of practical natural language grammars. In R. Studer, editor, *Natural language and logic: international scientific symposium*, Lecture Notes in Computer Science, pages 118–145. Springer Verlag, Berlin, 1990.

[21] K. McKeown and D. Radev. Generating summaries of multiple news articles. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 74–82, Seattle, Washington, 1995.

[22] C. D. Paice. Constructing literature abstracts by computer: techniques and prospects. *Information Processing and Management*, 26:171–186, 1990.

[23] K. Preston and S. Williams. Managing the information overload: new automatic summarization tools are good news for the hard-pressed executive. Physics in Business, 1994.

[24] L. Rau. Conceptual information extraction and information retrieval from natural language input. In *Proceedings of RIAO-88, Conference on User-oriented, Content-Based, Text and Image Handling*, pages 424–437, Cambridge, MA, 1988.

[25] L. Rau, R. Brandow, and K. Mitze. Domain-independent summarization of news. In *Summarizing Text for Intelligent Communications*, pages 71–75, Dagstuhl, Germany, 1994.

[26] G. Salton, A. Singhal, C. Buckley, and M. Mitra. Automatic text decomposition using text segments and text themes. In *Seventh ACM Conference on Hypertext*, Washington, D.C., 1996.

[27] K. Spärck Jones. Discourse modelling for automatic text summarising. Technical Report 290, University of Cambridge Computer Laboratory, Cambridge, England, 1993.

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

[29] J. Tait. *Automatic summarising of English texts*. PhD thesis, University of Cambridge Computer Laboratory, Cambridge, England, 1983. Technical Report 47.