Automatic summarising: the state of the art

Karen Spärck Jones
Computer Laboratory, University of Cambridge
William Gates Building, JJ Thomson Avenue, Cambridge CB3 0FD, UK
sparckjones@cl.cam.ac.uk

This paper in its final form will appear in *Information Processing and Management*, Special Issue on Automatic Summarising, 2007.

Abstract

This paper reviews research on automatic summarising in the last decade. This work has grown, stimulated by technology and by evaluation programmes. The paper uses several frameworks to organise the review, for summarising itself, for the factors affecting summarising, for systems, and for evaluation.

The review examines the evaluation strategies applied to summarising, the issues they raise, and the major programmes. It considers the input, purpose and output factors investigated in recent summarising research, and discusses the classes of strategy, extractive and non-extractive, that have been explored, illustrating the range of systems built.

The conclusions drawn are that automatic summarisation has made valuable progress, with useful applications, better evaluation, and more task understanding. But summarising systems are still poorly motivated in relation to the factors affecting them, and evaluation needs taking much further to engage with the purposes summaries are intended to serve and the contexts in which they are used.

Automatic summarising: the state of the art

1 Introduction

In the last decade there has been a surge of interest in automatic summarising. This paper reviews the state of the art for this challenging natural language information processing (NLIP) task. There has been some progress, but there is much to do.

The rest of the introduction notes the stimuli for this development. Section 2 provides a framework for the subsequent review, and Section 3 considers summary evaluation so far. Section 4 examines system coverage of the factors affecting summarising, and Section 5 system design types and examples. Section 6 assesses the overall state of knowledge and implementation. In general references are illustrative: the literature is now too large for exhaustive citation. For a much fuller discussion of the whole, see Sparck Jones (2007).

The Dagstuhl Seminar in 1993 (Endres-Niggemeyer et al. 1993) marked the beginning of a new research phase in summarising, and the 1997 ACL Workshop (*ACL-97*) that of the larger effort reflected in a succession of subsequent workshops, especially those associated with the DUC programme (*DUC*), and evaluation activities led by DUC and NTCIR (*NTCIR*).

Research on summarising since the mid-90s has been driven partly by statistical approaches going back to Luhn (1958), and partly by successes with combined symbolic and statistical approaches to other NLIP tasks like information extraction (IE) and question answering (QA). The performance levels reached with statistical and hybrid techniques in other areas have suggested they could also deliver useful summaries, and the NLP tools now available, e.g. for parsing, have made system building and task experiment easier.

The rapid growth of publicly accessible text, notably on the Web, has also stimulated work on summarising, and has pushed it to tackle types of input material, summarising purpose, and output style well beyond the classical summarising focus on journal papers. Thus online news has brought the challenge of summarising multiple documents with overlapping content; search engines have emphasised quick filtering for search hits as a use for summaries; and multi-window interfaces make it easy to display linked summary views of sources.

The pressure to summarise the mass of sources has helped to fund the evaluation programmes that have encouraged summarising work. It has equally, in conjunction with the interactive, multi-tasking that computing supports, led naturally to research designed to answer the question: can shallow NLIP techniques give us something that can do a good enough job for the user? This review considers what answer summarising research in the last decade gives us. What have we learnt about summarising needs and summarising technologies?

It must be emphasised that this review is *not* intended as a tutorial, and has somewhat different goals from such valuable earlier publications as Mani and Maybury (1999) and Mani (2001). As a state of the art review it is designed to consider the nature and results of the very extensive work on, and experience of, summary system evaluation since e.g. Mani (2001), though to motivate this analysis the review takes into account the large growth of summarising research since the mid 1990s. Thus the review approaches the status of summarising research first from the point of view of recent evaluation programmes and the factors affecting summarising that need to be taken into account in system evaluation. Then to complement this discussion, the review examines system strategies (for convenience using fairly conventional strategy classes) to see both how these strategies interpret a general model of the summarising process and what evidence there is for the strategies' effectiveness, insofar as the evaluations to date have stress tested them: it is in fact hard to make solid comparisons or draw general conclusions about correlations between task conditions and strategy choice.

2 Discussion framework

To provide a structure for the review of summarising work that follows I will make use of several frameworks, for summarising systems, for the factors that affect them, and for approaches to evaluating them. (The specific publications cited in this framework presentation are used because they provide concrete handles for the subsequent review, not as claims to exclusive originality.)

System structure

I shall start with the summary definition given in Sparck Jones (1999), with text as default input and output, as follows:

> *a reductive transformation of source text to summary text through content condensation by selection and/or generalisation on what is important in the source.*

Summarising is thus about both information and expression.

Sparck Jones (1999) assumes a tripartite processing model distinguishing three stages, as shown in Figure 1: source text *interpretation* to obtain a source representation, source representation *transformation* to summary representation, and summary text *generation* from the summary representation. Definition and model are deliberately broad enough to allow for summaries of many different kinds, for processing from the minimalist to the wholly radical, and for different distributions of effort across the stages. For present purposes the structure model, though general, is useful as a tool for characterising and comparing systems.

Summarising factors

Figure 1 refers only to summarising systems themselves. But these cannot be context free. It is essential, as discussed in Sparck Jones (1999) and developed in Sparck Jones (2001) to consider the task for which summarising is intended, i.e. the *setup* within which a summary is used. The design and evaluation of summarising systems have to be related to the three classes of *context factor* shown in Figure 2: *input* factors characterising the source material including style and units, *purpose* factors including intended use and audience, and *output factors* including reduction and format. In some cases purpose fully determines output, but more usually leaves specific choices open. System outputs and the mechanisms producing them cannot be compared without reference to input characteristics and purpose requirements; and proper evaluation requires adequate purpose specifications.

Evaluation elements and levels

There are many choices to be made in planning and conducting evaluations. Figure 3 applies the decompositional approach to evaluation presented in Sparck Jones and Galliers (1996) to an imaginary summarising situation, showing an evaluation *remit* and *design* in detail. The design is geared to the remit, and takes the input and purpose factors of Figure 2, along with determined output properties, as *environment variables*. The *system parameters* embody the processor structure of Figure 1. The evaluation performance criteria, data and measures follow from the remit.

NLIP system evaluation, including summary evaluation, has been applying the distinction between *intrinsic* and *extrinsic* evaluation. The former refers to the extent to which a system meets its own objectives, the latter to its functional effectiveness in context. This functional evaluation relates to the purpose factors of Figure 2 and has to consider the setup within which the system operates. Summaries may, for example, be intrinsically evaluated against a system objective of delivering well-formed discourse, or extrinsically evaluated against a setup requirement for summaries that can replace full scientific articles for information review by busy researchers.

However experience with summary evaluation suggests the intrinsic/extrinsic distinction is too crude and that a finer granularity is needed, from *semi-* through *quasi-* and *pseudo-* to *full-purpose* evaluation as shown in Figure 4. Evaluation without any reference to purpose is of extremely limited value, and putative intrinsic evaluations are in fact likely make tacit assumptions about purposes without recognising their implications.

The work done on summarising so far has been both limited in relation to the range of possibilities and very heterogeneous. It is also clear that what summaries are designed for, or de facto used for, matters. This suggests that starting the review which follows by considering the evaluations of the last decade can provide a useful context for the subsequent discussion of systems themselves, with the factors to which these have responded and the strategic

forms these responses have taken.

3 Summary evaluation

Earlier research on summarising included both informal single-system evaluation (Pollock and Zamora 1975) and more organised studies (Edmundson 1969; Earl 1971), comparisons between systems (Edmundson) and against baselines (Brandow et al. 1995). The growth of interest in summarising by the mid-90s prompted the SUMMAC cross-system evaluation (*SUMMAC* 1998, Mani et al. 2002). This was followed by the larger and sustained DUC programme (*DUC*). This, and its offshoots, has been valuable for the stimulus it has given to system building, for the information it has provided about systems and, perhaps most of all, for the way it has forced researchers to pay attention to the realities of evaluation for such a complex NLIP task. This includes issues about the interpretation and utility of concepts like intrinsic and extrinsic, and about the detailed design of task evaluations.

The DUC programme's initial road map envisaged a progression from internal systems-oriented evaluation to external purpose-oriented evaluation. But devising true task-oriented evaluations for summarising, i.e. ones that engage properly with the *contextual* task for which summaries are to be used, has turned out more difficult than for other NLIP tasks where evaluation programmes have been able to limit the evaluation scope without gross oversimplification. Thus for document retrieval evaluation has worked with relevance assessment alone, but this is a genuine core task need; for translation on the other hand, evaluation can be confined to text segment comparisons which are feasible and manifestly useful, if only indirectly, for task purposes. As the next subsection shows, summaries cannot be assessed in either of these styles, so attempts to apply evaluation methodologies used for other tasks to summarising may be misconceived.

Summary evaluation concepts

The problems of summary evaluation, and some common evaluation strategies, as reviewed in e.g. Mani (2001), already appear in Pollock and Zamora (1975). Much of the evaluation done in the last decade represents an attempt to firm, and scale, up evaluation, and to move from no, or at most presumptive, task evaluation to real task effectiveness testing.

Much earlier work followed Luhn (1958)'s *extractive* paradigm. There were therefore generally none of the problems about individual sentence well-formedness that could occur with *non-extractive* strategies. The main issues were about sentence choice, and about summary cohesion and coherence, i.e. about discourse well-formedness. However there is no good reason to limit summarising to extraction, so evaluation has to cover sentence well-formedness as well as discourse well-formedness. There is further no reason to limit summarising to straightforward source *reflection* in information and expression. Add both of these to the

primary challenges of capturing important content and condensing long source to short summary, and evaluation is bound to be hard. The evaluation strategies considered below represent attempts to make evaluation tractable by tackling its easier options first. But they have not always been well-founded.

Text quality

Summaries do not have to consist of running text: they may be phrase lists or tables with slot phrase fillers. But as running text is commonly required, it seems reasonable to begin evaluation by checking for 'proper' sentences and 'proper' discourse. NLP technology is now good enough to allow 'preliminary filtering' evaluation that checks summary sentences for specific syntactic properties like subject-verb agreement. This is a reasonable evaluation strategy for summarising in general even if individual applications will tolerate ill-formed output; and it has been one strand in DUC.

Quality questions are easiest to devise for local phenomena. It is harder to check global well-formedness beyond specific questions, e.g. about referents for anaphoric expressions, or rather broad ones, e.g. text cohesion. Establishing global coherence may require subject knowledge, and can be apparent rather than real, e.g. anaphoric referents are plausible but incorrect.

Unfortunately text quality is too weak to be a system discriminator (Marcu and Gerber (2001). More importantly, though text quality appears purpose-independent, it does in fact refer to system purpose. The system objective, to deliver well-formed phrases, sentences or discourse, is geared to what this output is for. This applies even if many different uses all need well-formed output. Calling text quality evaluation intrinsic evaluation obscures this important point. It is better referred to as the *semi-purpose* evaluation of Figure 4, with the purpose made explicit in the evaluation design.

Concept capture

The first question about a summary, Is it alright as discourse? is thus not as straightforward as it looks. The second, Does it capture the key concepts in the source? is clearly much less straightforward. Even for a reflective summary, it depends on being able to identify the key concepts in the source and, since these are normally complex relational concepts, recognising that, whatever condensing transformation the summarising process has involved, they figure in the summary. Though some applications may be highly specific about what constitutes key source material, most cannot be. Thus direct evaluation of content capture implies source markup for important content and summary inspection to see it is there.

This is a matter of human judgement, and obviously hard to control. Even for the simplest version, where humans are asked to mark up source sentences to extract, they do not agree (Rath et al. 1961). Using multiple judges may deliver more- and less-agreed sentences, but summarising is not just extracting source sentences. The underlying problem is that the instruction 'mark impor-

tant source content' is unavoidably vague. This applies even when a need for reflective summaries is taken for granted, and reflective summaries are not always wanted. Professional abstracters (Rowley 1982; Endres-Niggemeyer 1998) markup sources, but within a complex process that cannot be taken over to support evaluation as direct source-summary comparison.

Using multiple markers may reduce variation, but is extremely expensive. It is therefore natural to look for other ways of showing that summaries have captured important (appropriate) source content. Edmundson (1969)'s and Brandow et al. (1995)'s requirement for summary acceptability against source was very weak. Using questions that can be answered with the source and should be answerable with the summary looks like a strategy with more leverage. Morris et al. (1992) used educational reading comprehension questions, and Minel et al. (1997) and Teufel (2001) use more sophisticated questions about argument structure; SUMMAC (*SUMMAC* 1998, Mani et al. 2002) used questions about significant source content that should be answerable from summaries, and Kolluru and Gotoh (2005) argue the method is robust against human subjectivity.

But with rich sources the range of possible questions is enormous. More importantly, as Minel et al. point out, this strategy again involves an implicit reference to summary context and purpose. Farzinder and Lapalme (2005) address this point by using lawyer-oriented questions, but still have to justify the specific questions used. Question answering is appealing as an apparently focused, low-cost way of evaluating summaries by comparisons with sources. But it is in fact a form of *quasi-purpose* evaluation, and is methodologically unsound when divorced from reference to, and control by, knowledge of summary purpose that can mandate appropriate questions.

Gold standards

These difficulties, of pertinence, informativeness and cost, in working with direct source-summary comparison, as well as its own obvious attractions, has encouraged system summary evaluation against human reference, model, or *gold-standard* summaries. Humans know how to summarise and can therefore be relied on to capture important source content, and to produce well-formed output text. Summary-summary comparison has the particular advantage, compared with source-summary comparison, of removing the explicit condensation requirement. The strategy suits extractive summarising, where whole sentences are compared and it can be applied automatically. But it can be developed to cover content *nugget* comparisons, with human assessors to mark up nuggets and to judge their similarity, but within the restricted limits of short texts. The SEE program used in DUC (Over and Yen 2004) provides a tool to support this.

But as Rath et al. (1961)'s study implied, humans do not extract the same source sentences to give a single gold standard; and using multiple gold standards brings new problems. Thus as McKeown at al. (2001), Daumé and Marcu (2004) and Harman and Over (2004) point out, model summary variations may swamp system differences, so model-system differences have no clear implications for summary, especially system summary, value. Comparing mul-

tiple models to establish relative agreement on sentence or nugget choices, as in the Pyramid scheme (Passonneau et al. 2005) appears to offer a way out. But with nuggets this substantially increases the evaluation effort. Moreover, as Daumé and Marcu demonstrate, even with very constrained summarising specifications, humans produce different summaries, and van Halteren and Teufel (2003) show that to counteract model variation many reference summaries are needed.

Gold-standard evaluation has the apparent advantage that deference to summary purpose can be built in. If the human summaries are designed for purpose, comparing systems summaries with the models will take account of system fitness for purpose. But when pre-existing summaries are used, the purposes they were intended to serve may be unknown. Further, even where the human summary purpose is known, there may be no evidence it meets it well. Yet further, where system summaries differ from human ones, this need not mean they are less fit for purpose.

Gold-standard comparison has nevertheless been seen as sufficiently attractive, and operationally viable, for it to have been widely used in both evaluation programmes and individual tests during the last decade. For extractive summarising in particular, it has been developed in the ROUGE program (*ROUGE*) to allow, by using ngram rather than sentence comparison, for similar but not identical sentences, for phrasal summaries, etc., and to factor in as many model summaries as are available. Indeed where model summaries are not available it can compare a system summary against a set of other system summaries, Lin (2004) shows it correlates reasonably with human coverage judgements, though variably with summary specifications.

ROUGE may thus have some utility as an indirect, but automatic, apparatus for providing some performance data on system summaries. But it has been primarily applied to extractive summarising and is an intrinsically coarse measure providing little if any diagnostic information. The problems of model summary variation remain and further, as Daumé and Marcu, and Harman and Over, point out, of variation among human assessors. This applies wherever absolute or comparative judgements are required e.g. Does summary S answer question Q? Is nugget N1 the same concept as N2? and also, for any mode of evaluation depending on concept-based markup, on deciding that some text segment S constitutes a nugget. The implication is that multiple measures of summary performance are needed, especially since, as McKeown et al. (2001) show, they rank systems differently, and that wherever human judges are required, measures of inter-judge agreement should be applied.

Since simple extractive summaries suffer from lack of text cohesion or coherence, and ROUGE does not address summary discourse characteristics, some proposals have been made for forms of automatic model comparison that address cohesion (Hori et al. 2004) and coherence, as in Santos et al. (2004)'s graph structures. At a higher level, Amigo et al. (2005) put forward a more ambitious gold-standard methodology than those discussed so far, that uses probabilistic techniques to assess, choose among, or combine, different similarity metrics for source-summary comparison. But all the gold-standard strategies depend on

having satisfactory model summaries (or competing equally satisfactory ones). The gold-standard approach is thus properly quasi-purpose evaluation rather than pure intrinsic evaluation. It therefore requires an adequate characterisation of the purpose that summaries, including the human models, are intended to serve, and some evidence the models do this.

The foregoing implies there are early limits to what can be learnt about the merits of summarising systems without reference to summary purpose. However as proper purpose-driven evaluation is difficult and expensive, as the illustrations in Sparck Jones (2001) and Sparck Jones and Galliers (1996) imply, simpler and cheaper substitutes are a natural first choice. Given available data and current test conventions, system builders have been tempted to pursue a 'suck it and see' approach to automatic summarising. But well-justified gold standards are a clear requirement even for this limited form of evaluation. It also has to be done on an adequate scale. DUC and other programmes have increased test scale, but it is still fairly modest and, as Jing et al. (1998) show, summary evaluations may be very sensitive to data and context conditions like summary length. The range of environment variables and system parameters covered in tests has slowly increased, but sensitivity analysis is still too rare.

Baselines and benchmarks

Gold standard summaries have been taken as setting a target performance level for system summaries. This target performance does not define a task upper bound and, as with retrieval, meaningful upper bounds are relative to specific ground conditions. They may still be useful, and Lin and Hovy (2003) suggest that it is possible to determine upper bounds for particular summarising strategies.

It has, on the other hand, become common to set *baselines* for summary performance. One, for extractive summarising, has been *random* sentences. A more sensible one for news, used in Brandow et al. (1995) and adopted in DUC and elsewhere, has been *lead* sentence (or n words). This particular baseline may not apply elsewhere but having some simple baseline is a valuable check on system performance. It has also become increasingly common, with extractive summarising along Luhnian lines and also taking account of retrieval experience, to work with the *benchmark* performance set by sentence selection based on some version of $tf * idf$ word weighting. It could be useful to establish some particular way of doing this, for both single- and multi-document summarising, as a community benchmark.

Recognising purpose

The need to cross the old intrinsic/extrinsic boundary and address summary purpose more directly is clear. The discussion so far implies that even with more limited approaches to summary evaluation, it is helpful to place them within the framework of a remit and design analysis like that shown in Figure 3. Taking purpose seriously makes it essential. Figure 5 illustrates a purpose-

based summary evaluation in detail, showing the choices made in remit and design, along with some variations.

This wombat example may look frivolous, but has real summarising analogues. Thus police reports were inputs for alerting summaries in the POETIC project (Evans et al. 1995), and other parallels, for example potential database entries as outputs or questionnaires as evaluation mechanism, appear below. The illustrations show that though both alerting evaluations are extrinsic, they differ in purpose relationship. Questionnaire-based evaluation is a form of *pseudo-purpose* evaluation: the audience is real but the questions address driving context and behaviour. The police accident-data evaluation, suitably before- and after-detailed, would be a *full-purpose* evaluation.

Both cases assume the same output summary form. But there are plausible alternatives for the alerting example, which would also need assessing, again raising issues about how to do it. POETIC, as noted, produced alerting summaries, but with heavy contextual constraints on their factual character and timing. This could make them easier to assess for effectiveness than the wombat case, but there were still possible alternatives to evaluate.

The ramifications of context are well-exhibited by *BMJ* (the *British Medical Journal*) summaries. Editorials and news items are summarised by lead-text extracts; but research papers have formatted abstracts with headings subsuming separate mini-abstracts, which may be phrasal or telegraphic, as shown in Figure 6. These differences partly reflect source ones, but much more obviously purpose ones, with the abstracts designed to meet readers' interests in clearly and conveniently presented test results. Even here different readers may have different particular interests (e.g more on an old topic, a new topic), that the abstracting policy does not explicitly address.

Purpose evaluations

The DUC programme has had proper purpose evaluation as its goal. The need to address the task for which summaries are intended has long been recognised (e.g. Hand 1997) if only because, as Okurowski et al. (2000) make clear, real world situations introduce desiderata and complexities that make focusing on summarising systems alone dangerous or wasted effort. Thus in spite of the challenges that purpose evaluation presents, there have been some purpose evaluations in the last decade, albeit pseudo- rather than full-purpose ones.

The main summary use considered so far has been for relevance filtering in retrieval. This was assumed in Pollock and Zamora (1975), and tested in Brandow et al. (1995), Mani and Bloedorn (1997) and Jing et al. (1998), in SUMMAC (*SUMMAC* 1998, Mani et al. 2002), and by Wasson (2002). These evaluations properly compared assessments on summaries against those on sources. Dorr et al. (2005) compared assessments on summaries with reference assessments of sources, i.e. gold-standard *annotations* on sources, which changes the people involved without proper control. Earlier tests used generic, not task-oriented, summaries, but Tombros et al. (1998) showed query-biased summaries worked better than pre-existing generic ones.

There have been enough relevance-filtering evaluations to suggest that, even if the evaluations have been far from perfect, automatic summaries are good enough for this use (and, by implication, for varied input and other purpose factors like audience). But it is also the case that very different summaries are equally effective, because this is not an exigent use. Effective performance for this task cannot therefore be taken as an indicator for other tasks and retrieval as a proxy evaluator for these. The evidence so far, moreover, only supports the generalisation that automatic summaries suffice for generic topic relevance, just as the query-oriented snippet summaries Web engines now offer do. Relevance filtering may be subject to more specific conditions that might be more taxing or discriminating.

Retrieval has established test protocols for core functionality. Other summary uses lack these. However some task evaluations have been done, both for tasks related to retrieval and others. Related tasks include support for browsing, and question answering. Browsing is hard to evaluate and Miike et al. (1994) report only simple time-based evaluation. Hirao et al. (2001) assessed summaries for efficiency in supporting question answering. Summarising has also been used to improve indexing and matching within retrieval systems (Strzalkowski et al. 1998; Sakai and Sparck Jones 2001; Lam-Adelsina and Jones 2001; Wasson 2002); but evaluation here uses standard retrieval methods.

The other major use evaluated so far has been report generation, as digests or for briefing. Minel et al. (1997) evaluated summaries as potential support for writing source syntheses. McKeown et al (1998) report an informal user-oriented study of the value of patient-oriented medical literature summaries, Jordan et al. (2004) a more substantive evaluation of data-derived briefings in a clinical setting. In evaluating Newsblaster (*NWBL*), McKeown at al. (2005) used report writing as a means of evaluating summaries as sources of facts: the reports were not the primary evaluation subject.

All of these tests constitute useful attempts to tackle summary purpose. But even the most solid are still pseudo- rather than full-purpose evaluations.

Summary evaluation programmes

The DUC evaluations

DUC has been the first sustained evaluation programme for automatic summarising. It is exhibited in detail in its workshop proceedings and analysed in the test cycle overviews (*DUC*). I shall summarise it here to see what it has shown about summaries, summarising strategies, and system effectiveness.

The programme was based on a broad road map (Road Map 1) that envisaged a gradual progress from less to more demanding conditions on all dimensions, input, purpose, output and evaluation, for example from single document, reflective, extractive summaries to multi-document, transformative non-extractive summaries, from undemanding 'open' uses to demanding 'closed' ones, and from limited to full purpose evaluation (note that "task" within DUC has a more specific meaning than in this review). However changes soon had to

be made, to accommodate community interest in multi-document summarising, and the difficulties of designing and conducting evaluations. The programme has therefore had two phases, from DUC 2001 - DUC 2004 and, after a revision for Road Map 2, from DUC 2005 onwards. The main features of the programme are shown in Figure 7.

The first phase worked with news, with text quality and model comparison evaluation and, gradually, moved towards potential purpose considerations, for example by requiring event-oriented summaries and judging summaries for responsiveness to question topics, i.e. from semi- to pseudo-purpose evaluation. However it also introduced some more taxing constraints, notably summarising automatically translated sources. The participants explored a range of strategies, all essentially extractive but ranging from wholly statistical ones to combinations of statistical and symbolic techniques and applying hybrid combinations to all processing stages as well as to support resource construction. Participants sometimes used the same system for different DUC tasks, sometimes different ones, as with GISTexter (Harabagiu and Lacatusu 2002) and Lite-GISTexter (Lacatusu et al. 2003). However as Figure 7 shows, system performance up to 2004, while better than lead baselines, was inferior to model performance, especially for content coverage.

These results would not necessarily mean automatic summaries could not be useful for particular purposes. But the task references introduced in 2003 and 2004 were very undemanding, and were assessed by model comparisons, i.e. indirectly. The comparison methods used also presented problems, ROUGE through being uninformative and nugget (selection and) similarity judgements through subjective variation. The pseudo-purpose evaluation by responsiveness to questions was undermined by the lack of a working context.

The difficulties and costs of evaluation, evident by 2003, stimulated study in 2004 to see if ROUGE could suffice for coverage evaluation, without a need for the nugget method. But though the two methods correlate fairly well, not surprisingly given the emphasis on extractive summarising, ROUGE is too weak to use as a sole evaluator. It was also evident that the text quality questions were of limited value for extractive summaries, in effect acting only as a threshold.

The DUC programme's complexities (and ambitions) are apparent in the detail for DUC 2003 and 2004 shown in Figure 8. DUC has tried to advance summarising and summary evaluation in orderly tandem, but in practice has been driven to ad hoc changes that inhibit systematic progress assessment. Thus DUC 2004 tried to solve some problems by using ROUGE, but this introduced new ones; and Arabic sources changed another test element at the same time. The revised Road Map 2 sought more control with both a tighter focus and a more explicit purpose orientation, by requiring short, multi-document summaries geared to carefully-formed user topics with associated questions (see Figure 9), and using both ROUGE and nugget for coverage comparison as well as responsiveness assessment. The tighter focus has been an advantage and has been continued in DUC 2006. However, while the task may be more demanding for systems, relative performance in DUC 2005 was as before, with systems better than baselines but clearly inferior to the human models.

Overall the lessons from DUC, as manifest by DUC 2005, are that systems can produce summaries of different types, but may also produce the same type, to similar eventual purpose effect, by different means; and that evaluation continues to be barely tractable.

Other programmes

The second major programme has been the NTCIR one (*NTCIR*), over three Text Summarisation Challenge (TSC-1 - 3) cycles. The programme was similar to DUC but institutionalised the extract/abstract (non-extract) distinction; it also attacked task evaluation following SUMMAC models, i.e. used pseudo-purpose evaluation.

The programme covered both single- and multi-document summarising at different lengths, for Japanese; its intrinsic evaluations used semi-purpose evaluation by readability and 'degree of revision', as well as vocabulary and coverage comparisons with models as quasi-purpose evaluation. Its extrinsic evaluations were for summaries as relevance filters in TSC-1, compared with full sources, and with a modified question-answering in TSC-3, where summaries were simply checked as containing answer strings.

As with DUC, successive cycles developed the protocols; the most pertinent results, for TSC-3, were similar to DUC's with relatively low scores for both extracts and abstracts on coverage and similar measures, much below human models. The system summaries did better on the question answering than on coverage, but it is not very clear what this form of question answering implies.

The topic-with-questions test used for DUC 2005 and 2006 is closely related to one of the forms of question answering studied in the TREC QA evaluations (Voorhees 2005a, 2005b). So-called definition questions invited extended responses, not single factoids. However responses were specified as sets of nuggets, not coherent text. Questions in series were also investigated, with later responses supplementing earlier ones. Evaluation depended on assessors identifying appropriate nuggets (perhaps retrospectively using system results) and on specifying some as vital.

These response summaries did not, however, start from particular sources: material could be taken from anywhere without regard for source content as a whole or source relationships. As summaries, these QA responses are wholly selective, not condensing. The evaluation has also been narrowly gold-standard, without regard for task context except to assume that in reality, users will recognise useful responses. The TREC QA evaluations have supplied valuable experience in focused text analysis, but have so far been limited if viewed as addressing one legitimate form of summarising.

Assessment of evaluations

We can see that, along with some definite, if modest, progress in building summarising systems, evaluation is more complex than it appeared to be. It is evident that it is necessary to recognise how crucial the task context is and how

dangerous the idea of intrinsic, gold-standard evaluation can be. It is generally impossible, even with fine-grained methods like nugget comparison, to predict summary utility. At the same time it is hard for developers, working with complex systems, to apply even such detailed data to improve system designs. The Catch-22 situations is clear in Lin and Hovy (2003): they attribute poor system performance to human gold-standard disagreement, so humans should agree more. But attempting to specify summary requirements so as to achieve this may be as much misconceived as impossible.

Outside the programmes summary evaluation is increasing but is primarily intrinsic. So there is little solid information about what makes summaries work in contexts. Radev et al. (2003) report a substantial comparative evaluation across multiple systems, with multiple measures, against gold standards and in a limited retrieval task. The former showed that measures that factor in inter-judge agreement are more satisfactory, but different measures rate systems differently, so is difficult to draw significant inferences about fitness for purpose from any of the results.

Operational summarising systems have nevertheless appeared in the last decade, within Web search engines, or freestanding like Microsoft's Summariser and Newsblaster (*NWBL*). The presumption has to be that users find the results helpful. but there is no information about this for the industry systems and McKeown et al. (2005)'s Newsblaster evaluation was limited in scope. All these operational systems are general-purpose, even the Web engine ones within the generic search situation, so they would be very hard to evaluate overall. In practice also, things are different. Moens and Dumortier (2000)'s automatic summaries for magazine articles were designed to prompt magazine purchases. A proper evaluation would check whether sales increased. But in fact the commercial publisher was sufficiently impressed by informal comparisons between the authors' summaries and his existing system ones to just install the authors' system.

4 Factors explored

In Section 2 I referred to the factors affecting summarising, as illustrated in Figure 2. This section considers the extent to which summarising systems in the last ten years have explored these factors, e.g. the types of source material. This covers both cases where systems have had to respond to specific factor values e.g. source medium or have explicitly chosen to respond to them, e.g. the discourse properties of transcribed speech, and cases where systems have not been designed to respond but their outputs and performance may nevertheless have been affected by factor values, e.g. source genre. The factor-by-factor review which follows, with illustrative references, shows that factor coverage has been grown over the decade, but is still patchy and limited.

Note that input, purpose, and output factors are not simply mapped on to interpretation, transformation, and generation stages. Thus output factors, for example, may determine transformation operations, not generation ones, that

eventually deliver outputs with particular properties.

Input factors

First, input factors, as listed in Figure 2.

Form factors

Form factors refers to the subclass of specific factors list below. These source form factors are the most important input factors. The dominant input type in summarising research has been news material from agency text streams. It is available, does not require technical domain knowledge, and is of interest to many potential system users. It has featured largely in evaluation programmes like DUC, stimulating further comparative experiment with the programme test data. However other types have figured, as indicated below, e.g. legal material.

Language

Considering specific form factors, English has been the main language (e.g. in DUC), with substantial effort in Japanese (see *NTCIR*) and work on Chinese (Chan et al. 2000), Dutch (Moens and Dumortier 2000) and German (Reithinger et al. 2000), and both raw Arabic (Douzidia and Lapalme 2204) and automatically-translated Arabic news in DUC 2004. Systems that deploy NLP resources like dictionaries have to respond to language, but statistical systems need not or do it in minimal ways as in stemming.

Register

Register here refers to linguistic style, as in popular, scholarly etc, that in principle needs response in summarising. News taken as instantiating the popular register has clearly figured, but so have other registers: technical articles as scholarly (Saggion and Lapalme 2000, 2002; Teufel 2001, Teufel and Moens 2002), legalese (Grover et al. 2003), email (Corston-Oliver et al. 2004), technical chat (Zhou and Hovy 2005). More interesting registers include lectures (Nobata et al. 2003) and presentations (Furui 2005), dialogues (Zechner 2001, 2002) and meetings (Murray et al. 2005). But register has not generally been recognised as a processing condition (though statistical methods implicitly respond to it), except for speech where e.g. Zechner responds to restarts and Furui cleans up 'untidy' transcriptions.

Medium

Most sources have been text, some text from speech. Input material may also include images and graphics. I am excluding image-to-image or graphics-to-graphics here (Futrelle 2004; Rother et al. 2006), interesting though they are. But sources combining language and image may call for combined summaries (Christel et al. 2002; Papernick and Hauptmann 2005), and combined sources need joint interpretation even for text summaries (Carberry et al. 2004). Non-text material, e.g. tables, has been taken as input for output text summaries

(Maybury 1995; McKeown et al. 1995; Jordan et al. 2004; Yu et al. in press). Image or data input clearly needs explicit system responses and Zechner exploits speaker separation for speech.

Structure

Structure refers to such marked 'external' structure as headings, boxes, rather than internal discourse structure e.g. repetition (though the distinction is not absolute). News stories have little explicit structure beyond top headings, and this has not usually been exploited though Moens and Dumortier (2000) pick up quotes as a marked form of structure. Teufel (2001) and Teufel and Moens (2002) use citations, which resemble quotes, in technical text. Other explicit structures signalled by e.g. headings, include those in legal material (Moens et al. 1997; Grover et al. 2003; Farzinder and Lapalme 2004), magazine articles (Moens and Dumortier 2000), and technical articles (McKeown et al. 1998, Elhadad and McKeown 2001; Saggion and Lapalme 2000, 2002). Web pages, source in Radev et al. (2001a), have complex structures with items that may need excluding (Berger and Mittal 2000b). Sato and Sato (1997) exploit Usenet threads as explicit structures linking multiple items.

Genre

News stories represents a genre that may be labelled reportage, mixing event and player descriptions. McKeown et al. (2002) respond to these in Newsblaster. Moens and Dumortier (2000) distinguish opinions and reportage for magazine pieces, Farzinder and Lapalme (2004) legal direction versus legal narrative. Narrative, description and argument are general genres, and McKeown et al. (1998) and Teufel (2001) and Teufel and Moens (2002) respond explicitly to argument and description. Salton et al. (1997) exploit the compact description that characterises encyclopedia article, Sato and Sato (1998) and Zechner (2001, 2002) the question-answer form of instruction and Reithinger et al. (2000) negotiation.

Length

News stories are usually short, technical articles long. Compression at 30% is plausible for news summaries, but not articles. Summarising for long sources has generally finessed length by worked with selected sections, e.g. Results, or by delivering only brief headline-type summaries or query-oriented snippets, though Nakao (2000) addresses book summarising. Multi-document summarising may have large input sets.

Other input factors

There are a number of other input factors, as follows.

Subject

News is varied in subject, but not normally opaquely technical, so systems do not require a subject infrastructure. Subject knowledge for the medical do-

main is used in McKeown et al. (1998) and for computing in Hahn and Reimer (1999), but statistical summarising, even for technical subjects, does not need subject resources. However there has been little subject-specific summarising research, other than for law (Wasson 2004).

Units

Summaries are normally for single sources, but working with news has stimulated summarising over source sets, i.e. multi-document summarising. With news streams, stories on the same topic may overlap heavily, more than in the linked passages that Salton et al. (1997) studied. Multi-document summarising therefore typically involves source clustering by topic or sub-topic and procedures to avoid content redundancy in summaries. Multi-document summarising has also been applied to Web pages (Radev et al. 2001a) and to sub-source units of the same type (e.g. results sections) in medical papers (McKeown et al. 1998, Elhadad and McKeown 2001).

Authorship

Authorship is not an obvious factor for news, and does not seem to have been considered except as, possibly, accounting for conflicting content in sources in multi-document summarising.

Header

The header factor refers to metadata assigned to sources rather than internal properties, e.g. assigned indexing Keywords, though the distinction between header and structure factors is a loose one. Story dates attached to news stories may be used to order extracted material. Zhang et al. (2003) explore source reader annotations like highlighting and Sun et al. (2005) clickthrough data for Web pages.

Source properties clearly have implications for summarising. But research so far throws little light, except where summary purposes clearly justify explicit responses, on the gains to be made from heavy tailoring to the source properties listed, as opposed to light system tuning or the responsiveness, at least to some properties, that statistical approaches automatically achieve. This is primarily because effort has focused on general-purpose approaches like statistical extraction, but also because it is not obvious how some properties can be identified or used. Linguistic characteristics at the primary text level are a different matter, since summarising systems are designed to respond to vocabulary, types of syntactic unit, etc.

Purpose factors

Use

The most important purpose factor is use, what a summary is for. This is the major influence on summary content and presentation (Sparck Jones 2001). There are a number of generic uses, some linked to sources like supporting source preview or source relevance filtering, others independent, like briefing or alerting. Generic uses can take many specific forms, uses may be combined, be prompted e.g. by queries, be for people or other systems.

This breadth and fluidity makes it difficult to apply intended uses to guide summarising and motivate evaluation. Uses need to characterised as fully and tightly as possible for system leverage and informative performance measurement.

Much of the summarising work done so far has not referred to summary use. This is partly because summarising is often reflective, so will by default meet the common requirement of summaries for any use, namely showing prominent source content; partly because system outputs may be designed to be multi-purpose; partly because users can themselves interpret summaries to apply them to their own uses; partly because system designers, e.g. of Web engines, may not be able to acquire appropriate information about users' purposes. Moreover as noted, gold-standard evaluation assumes use is taken into account via the model. Finally, except for specific applications, summarising systems so far have not obviously delivered good enough output to make developers move beyond the basics.

However as Section 3 showed, evaluation effort is moving to purpose-focused evaluation as a lever in system design and summary assessment. Evaluation by relevance filtering and the provision of Web engine snippet summaries also suggest that addressing purpose may actually, in some cases, show that apparently poor summaries are quite fit for purpose.

References to potential summary uses have often been extremely vague. However systems have been designed in the last decade that have been more definite about envisaged uses, even if they have not been sufficiently tested for purpose.

The main envisaged use for summaries has been in support for document retrieval, and in particular for relevance filtering. As noted earlier this dates from initial summarising research, has featured in evaluations in SUMMAC, DUC and NTCIR as well as, e.g. Brandow et al. (1995), Tombros et al. (1998) and Dorr et al. (2005). This use, narrowly defined, has been manageable and satisfiable one. Other related uses that have been taken as system motivators, though not necessarily with evaluation, have been summarising for 'skimming' source overviews (Strzalkowski et al. 1999; Boguraev and Kennedy 1999) or, more generally, browsing (Miike et al. 1994). Specialised versions of this use have been aimed at the disabled, in audio telegraphese for 'scanning' for the blind (Grefenstette 1998) and subtitle gisting for the deaf (Vandegehinste and Pan 2004). The implications of particular uses are shown in Moens and Dumortier (2000)'s highlighting summaries designed to stimulate purchases.

The other main type of task has been briefing, primarily in application-specific forms e.g. medical literature summaries geared to patients (McKeown

et al. 1998, Jordan et al. 2004), 'to-do' lists (Corston-Oliver et al. 2004), but also in more open form (Mani et al. 2000), though not all have been evaluated. Definition question answering in TREC (Voorhees 2005b) is a form of briefing. Support for report writing, envisaged in Minel et al. (1997) and used to evaluate Newsblaster (McKeown et al. 2005), is similar; and the topic/question-oriented summarising tested in DUC and Hirao et al. (2001) can be seen either as briefing or reporting per se, or as support for these.

In these cases use has influenced, or at any rate justified, system design. In others, uses have been retrofitted, possibly only indirectly. Thus as manual headlines are common and have assumed uses as 'one-line' summaries, generating headline summaries is taken as a system goal (Dorr et al. 2005) without reference to any particular use. More generally systems are built using plausible generic ideas and available technology, subsequently tested for some use, and modified as required. This is the opposite of a sound design process.

Audience

In general professional abstract writers for academic paper sources, just as authors who write abstracts, assume a technically-informed audience like the one for the sources. But summarising 'interesting' science papers for newspapers assumes a different type of audience. Executive summaries allow for audiences that differ from full report readers.

News material has (at least) two audiences: 'ordinary' readers, and analysts. Though these two classes may overlap in specialist knowledge and focus, they are essentially distinct. Systems like Newsblaster assume ordinary readers; the DUC programme, especially in moving up the purpose evaluation levels, has increasingly assumed summaries are for analysts. However these differences in audience type have not been factored into system design. Radev et al. (2001a) take a wide and varied audience for granted in Web page summarising, and Web search engines have to do this with their snippet summaries.

Other summarising research has assumed narrower types of audience, for example legal professionals in Moens et al. (1997) and Farzinder and Lapalme (2004), academics in Teufel (2001) and Teufel and Moens (2002). Corston-Oliver's briefing summaries are for the members of an organisation, McKeown et al (1998) and Jordan et al. (2004)'s are for the doctors in a particular community. Reithinger et al. (2000)'s Verbmobil summaries have a different audience view since these are for dialogue participants responsible for the original long sources.

Some of these specific audiences have conditioned system design, e.g. in the medical case, or follow particular uses, like 'to-do' briefing. But in other cases audience type seems to follow straightforwardly from the nature of the source. In evaluation, broad audiences have often been represented by rough proxies, but specific ones need to involve their intended audience or very near proxies.

Envelope factors

Envelope factors refers to a subclass of purpose factors that covers other purpose conditions, e.g. summary locations, as follows.

Time

Time has not normally been a critical factor in summarising, or addressed in research. However Evans et al. (1995)'s traffic alerts had to be timely, Newsblaster has to roll forward with developing story lines, and Corston-Oliver et al. (2004)'s 'to-do' lists have to be delivered soon. Query-oriented summaries have normally to be returned promptly, as Web search engines show. Radev et al. (2001a) is unusual as a research-based paper that considers time from the engineering point of view, in scaling up for more users.

Location

Location has not been examined in detail, but is pertinent for any digital output to a workstation or other device because of the display, connection and interaction opportunities this allows, e.g. summary phrase highlighting (Boguraev et al. 1999) and other forms of visualisation (Corston-Oliver 2004), image linking (Newsblaster; Mani et al. 2000), clickthrough to related items, and user 'personalisation' through different views (Aone et al. 1997). While the summarising system itself may offer a range of facilities, summarising is increasingly embedded in a richer multi-function system environment, as in WebInEssence (Radev et al. 2001a) and MiTAP (Damianos et al. 2001).

Formality

Formality refers to specific requirements that are not deducible from use or audience, and often conventional in kind. It includes, e.g., legal constraints to avoid liability, author attributions for summaries. Using specific category headings for summaries as in *BMJ* (Figure 6) is on the formality border. Summarising research has concentrated so much on summary content it has largely ignored formality issues.

Triggering

Triggering clearly occurs for summaries produced in response to search queries. It is also natural, and possibly more complex, for alerting summaries, as in Evans et al. (1995) where each system input triggers an explicit decision on whether a new or revised alert is needed.

Destination

The default destination for summaries has been the human end user, assumed capable of interpreting the supplied summary appropriately. This applies not only to e.g. DUC-style summary quality assessment, but also to the nature of the context task. Destination conditions may interact with summarising strategies, for example where summaries are for a database subject to automatic query. There has been little work on destination implications, though the danger of misleading discourse connectivity in extractive summarising in particular has been recognised. Summaries as input to further system modules

in retrieval systems (e.g. Lam-Adelsina and Jones 2001) are not subject to very detailed destination constraints, but input to a translation module (Douzidia and Lapalme 2004) could guide syntactic choices to make translation easier.

There are many factors and factor values that affect summarising and thus in principle may influence system design. Few have so far been explicitly addressed, though the particular strategy adopted in Lite-GISTexter (Lacatusu et al. 2005), for example, was specifically motivated by the requirement for question-directed summarising. This poor factor coverage makes it hard to show there are types of strategy suited to types of purpose, and indeed individual purposes vary so much it may be impossible to generalise for purposes beyond such weak claims that for purpose type P, strategies of type T are not without utility.

4.1 Output factors

Input and purpose factors constrain output, but do not determine it, especially in the fine grain, e.g. in sentence syntax choices. In many cases they leave larger choices, e.g. between formatted output and running text, open until purpose-based evaluation establishes relative merits. Most automatic summarising has produced running text, whether extracted or generated, as a natural default in natural language use, but phrasal summaries may suit some purposes like skimming or retrieval assessment (Oka and Ueda 2000), and e.g. Zechner (2002)'s DIALSUMM can produce both for spoken dialogues. But summarising research has generally ignored many output factors, while in others one type of option has been selected as appropriate but without examination of finer sub-choices, e.g. using formatted output but without examination of layout details. Output factor choices apply to both transformation and generation stages.

Material factors

Material factors covers a subclass of factors that refer, broadly, to the relation between source material and summary material, as detailed below.

Coverage

Summary coverage of the source may be comprehensive or selective. Reflective summaries are comprehensive, while query- or topic-oriented summaries (as in DUC) are the main form of selective summary investigated so far. This applies to both single- and multi-document summarising. Heading-based summaries as in *BMJ*, McKeown et al. (1998) and Moens et al. (1997), and briefing and alerting as in Corston-Oliver et al. (2004) and Evans et al. (1995), are also selective.

Reduction

Reduction (sometimes called compression), defined mechanistically, has figured conspicuously in summarising research. Thus several DUC cycles called for

summaries at $X\%$ source length. This is a plausible way of controlling summary length with e.g. a ranked list of source sentences as candidates for summary inclusion. But some lengths, e.g. 30%, are only sensible for short sources. In general, reduction factors have been set more as system tests, as in Grewal et al. (2003), than because uses or users demand them. The other length specification in recent research has been the brief headline, as in DUC, Banko et al. (2000), Dorr et al. (2003) and Zhou and Hovy (2004), again without clear use requirement. Purpose has however been the motivation for local reduction through telegraphese or heavy global reduction e.g. for audio skimming (Grefenstette 1998), or to fit summaries in handheld devices (Boguraev et al. 2001; Corston-Oliver et al. 2001).

Derivation

Derivation refers to whether summaries reproduce source expressions e.g. as phrases or sentences, or express source content quite differently. NTCIR made derivation an experimental parameter as explicit extractive summarising, but properly it follows from use, e.g. Web search engine snippet summaries deliberately reproduce source text. In other cases use implies no replication constraints, as in Corston-Oliver et al. (2004)'s 'to-do' lists. Apparent intermediate cases, where source material is tweaked or truncated, say by sub-clause omission (e.g. Harabagiu and Lacatusu 2002), are essentially still replicative; but complex cases where source quotes may be lifted but placed within new text have not been investigated. Most systems deliver extracted text in some form, but this is because the systems are easier to build and may serve purposes sufficiently well, not as conscious response to purpose.

Speciality

Some audiences may require some reduction in specialised technical language and detail. This has apparently not figured in summarising research.

Style

Summary style, as a separate factor, refers to a loose but recognised notion, with informative versus indicative summaries the best-known examples. In human summarising the need is often for informative summaries (e.g. *BMJ*, Figure 6). Summaries using information extraction approaches are normally intended to be informative, e.g. for medical briefings (McKeown et al. 1998) or dialogue interactions (Reithinger et al. 2000), and so are summaries that respond to questions, as in DUC 2004. However summaries for source skimming (Boguraev and Kennedy 1999), or relevance assessment (Oka and Ueda 2000), are intended to be indicative. Saggion and Lapalme (2002) offer both, with informative summaries amplifying indicative ones.

Format factors

Aa the final subclass of output factors, along with output material sub-

factors and style, there are format sub-factors that purposes may mandate, but often do not.

Language

Output language is normally a deliberate choice, usually the same as the source language but sometimes different as in the DUC tests for English summaries for Arabic sources. Most summaries have been in English, with Japanese in the NTCIR programme, but there has been some work with other languages, e.g. German (Reithinger et al. 2000).

Register

Purpose may require a particular choice of language register or linguistic style. But most summarising research, because it is based on source extraction, reproduces the source register and thus makes the tacit assumption that this is also suitable for output. Where source text is tweaked this is to remove unwanted content, not change linguistic style. Phrasal summaries (Boguraev and Kennedy 1999), snippets and telegraphese, however, may be viewed as changing register as appropriate to summary purpose, which is quite explicit in Grefenstette (1998)'s audio telegraphese. This also applies to compressed headline summaries (Witbrock and Mittal 1999). Sato and Sato (1998)'s source rewriting to make summary answers to questions easier to understand in an instruction context is a deliberate register choice. Summary generation from deep source representations normally changes register by default.

Medium

Summarising research has explored non-text image output, both free-standing (Rother et al. 2006) and alongside text as illustrations (Merlino and Maybury 1999; Newsblaster), or with text as video annotation (Papernick and Hauptmann 2005). Grefenstette (1998) and Carberry et al. (2004) have audio output for disabled users. However there has been little work exploring alternative media options or combinations for relative task effectiveness, apart from Merlino and Maybury's. Taking a broader view of media, modern output devices offer a variety of presentation devices including visualisation (Boguraev and Kennedy 1999) and interaction opportunities (Aone et al. 1997; Ando et al. 2000; Mani et al. 2000; Radev et al. 2001a), though these have generally been offered as attractive possibilities rather than evaluated.

Structure

Purpose may specify structure, but there are often open alternatives, e.g. phrases in source order or alphabetical order. Summarising based on information extraction lends itself to output in a forms (Maynard et al. 2002; Mani et al. 2000) or tabular (Farzinder and Lapalme 2004) structure, while White and Cardie (2002) produce rich hypertext. The complex modes of visualisation mentioned under medium offer further structural possibilities, both for summaries themselves and their relations to other information entities (Radev et al. 2001a).

Genre

Research on text generation has explored genre, and genre choice is required in summarising from non-linguistic sources, as in Maybury (1995)'s choice of report mode in summarising source event data. Simple extractive summarising carries source genre over to summary by default, but with highly-selected material, especially in multi-document summarising, and where output sentences use but do not replicate input parses, output genre may be different from input. But, as in Newsblaster, and Elhadad and McKeown (2001), output genre may be a byproduct of the particular source content selection or treatment of different content types: thus in Newsblaster event and person summaries may naturally become narrative and description-oriented respectively.

4.2 Factor lessons

As the foregoing indicates, factor implications have not been comprehensively or systematically addressed in summarising research in the last decade. Except where attention is mandatory, or a response is beyond current technology, summarising work has been primarily directed to getting a system to deliver summaries that are not patently inadequate and may serve well enough for whatever summary purpose is in view.

This appears to have been the approach behind much extractive summarising, though it can also be more robustly justified as confirming Luhn (1958)'s belief that this simple strategy has value because it captures source properties that matter for summarising. DUC and other programmes, along with individual application projects have, however, encouraged more systematic exploration of the factor terrain though the challenges of system building and evaluation mean this has not progressed very far: as noted earlier, evaluation so far has not generally been very taxing, so researchers have been able to ignore demanding factor requirements or postpone comparisons between different, fine-grained output choices.

5 Systems: approaches and structures

Earlier work on summarising explored both shallow, essentially statistical approaches (Luhn 1958), deep symbolic aproaches (see Hahn and Reimer 1999), and hybrids (Earl 1970). More recent work, stimulated by evaluation programmes, test data, better language processing tools, and external task interests, has been far more extensive. It has pursued both generic approaches, though with more emphasis on statistical ones (e.g. NeATS, see Lin and Hovy 2002a; MEAD, see Radev et al. 2001b), and more development of systems combining statistical and symbolic techniques (e.g. SUMMARIST, Hovy and Lin 1999; Lite-GISTexter, Lacatusu et al. 2003). The relative emphasis on extractive approaches contrasts with earlier interests in text *meaning* representation and the role of discourse structure, as in Hahn (1990) and Endres-Niggemeyer et al. (1995), though these figure in current research.

It is impossible to review recent work in detail. I will therefore use the basic system structure of Figure 1 as a way of characterising systems as illustrated by cited examples, indicating the types of source and summary representation used and types of process involved, and then complement this analysis with brief accounts of exemplar systems. As in the previous section, I will consider whether or how far system approaches have been motivated by task factors. Indeed, as mentioned in the Introduction, this discussion of systems and their structure is not intended primarily as a review of recent and current systems for its own sake, but as a review that examines the structural possibilities that have been exploited for the light they throw on the relation between system structures and task requirements (though the limitations of evaluation to date make it impossible to draw strong conclusions about this). This review also, as mentioned, makes use of familiar system categories, and may also refer to systems covered in previous surveys, e.g. Mani (2001), but seeks to bring the analysis up to date.

For convenience, I will group systems as extractive and non-extractive. Each, the former especially, covers many variations and there is no absolute distinction between extractive and non-extractive. The recent interest in multi-document summarising, not considered in earlier research, complicates the picture but still falls within the Figure 1 model. The model is in any case a logical one. System implementations with their particular modules and process flows may look rather different, but it is helpful here to ignore their fine detail and use the general model for analysis and comparison.

Extractive strategies

Basic statistical approaches

It is natural to start with statistical approaches to summarising as these are simpler than symbolic ones and have been more widely pursued.

The simplest strategy follows from Luhn, scoring source sentences for their component word values as determined by $tf * idf$-type weights, ranking the sentences by score and selecting from the top until some summary length threshold is reached, and delivering the selected sentences in original source order as the summary. In this approach the actual sentences themselves are not part of the source or summary representations: these are just the source-ordered, or selected, sentence identifiers with their scores. Interpretation and generation map text to and from these, and transformation is essentially ranking and selecting them. Clearly, as characterisations of source and summary text meaning, the representations used are weak and indirect, with sentences treated independently and their meaning as a numerical function of their component word frequencies.

Treating sentences independently, however, means that summary sentences may repeat content. This can be dealt with by, e.g., applying Maximal Marginal Relevance (MMR - Carbonell and Goldstein 1998), so sentences are added to the

selection only if they differ from previous ones. But it also implies a richer source representation that records the actual words for sentences. In practice redundancy prevention may be done during generation so the summary representation also includes lexical data, but it is logically an element of transformation.

Multi-document summarising also requires richer processes and representations. Document sets with similar or related content have to be identified, whether within a whole file or as more tightly-related subgroups. These theme or topic groups are obtained by clustering using lexical data, and clusters are characterised by, e.g., centroid word vectors (Radev et al. 2000). Individual documents, or sentences, are then scored against the topic vectors. A cluster of documents on a broader topic is usually taken as the base for a single summary and the presumption is that the summary takes account of subtopics, again statistically identified. The source representation is therefore primarily a set of sentence identifiers with their subtopic scores. However the subtopics themselves may have relative importance scores and, since the sentences within a subtopic are likely to overlap in content, the sentence representations record component words for future redundancy processing. Transformation for sentence ranking, selection and ordering, and generation to deliver source sentences, are similar to the single-document case, but factor in subtopic coverage, e.g. by a round-robin strategy, and avoid redundant sentences, e.g. by applying MMR.

Enriched statistical approaches: lexical units and features

The basic statistical approach is clearly applicable to, e.g., query-oriented summarising through query-term matching at some selection point, and to other units than sentences, e.g. text windows. More importantly, it is naturally extensible to a more sophisticated treatment of the lexical elements for which statistics are computed. This includes both the types of units chosen and differential weighting for unit types. It may also refer only to the interpretation stage, which delivers sentence scores and source representations as before, or to approaches which include units in the representations so they are available for later operations, as long as the eventual output is text extracted from the source, perhaps with modest tweaking. (The boundary with non-extractive approaches is where the internal representations are used for new text, though this boundary is fuzzy).

Thus one major research line has been to use more varied and elaborate lexically-based sentence features for score computation, The features themselves may be statistically determined, e.g. using recurrent ngrams rather then words, or statistically-motivated multi-word units like word pairs, or statistically-based word groups indicating generic concepts, like Lin and Hovy (2000)'s topic signatures. More directly linguistic, i.e. symbolically-grounded, tactics include using lexical resources characterising word senses and relations like WordNet, applying stemming or morphological operations to merge variant word forms, or applying current parsing technology to identify significant types of sentence constituent, for example noun groups, or dominant structures like main verbs and their arguments. These souped-up statistical approaches, which may also be ap-

plied to topic identification, are illustrated by Barzilay and Elhadad (1999) and Harabagiu and Lacatusu (2005) and by SUMMARIST (Hovy and Lin (1999) and Lite-GISTexter (Lacatusu et al. 2003). Modern shallow parsing technology, including part-of-speech tagging, is quite robust, and can be used to identify linguistically-significant multi-word lexical elements like named entities, or phrasal concepts like Filatova and Hatzivassiloglou (2004)'s 'atomic events'. Earl (1970) first seriously investigated these ideas, but recent research has taken them much further.

Specific lexical items with importance-signalling properties, e.g. "conclusion" in some domain literatures have also been investigated (Teufel and Moens 1997, 2002), and so have other unit types with language-like properties, like Web links and URLs (Chakrabarti et al. 2001). Again, there are other forms of information including metadata that may be used as features for unit weighting, like title occurrence or typographical emphasis for words, and like paragraph initial position for sentences. Again, while Edmundson (1969) early on investigated these forms of information bearing on word or sentence importance, they are much more easily studied now.

Many of the systems built in the last decade have been based on the sentence extraction model. However one significant variation is where units like phrases (whether statistically or symbolically obtained) have replaced them. These are treated much like sentences, though they may require an additional step to choose which of many variant forms is output, and the summary itself is just a phrase list. Phrase list summaries, though very minimal, may be suited to tasks like relevance filtering or browsing in retrieval (Witten et al. 2000).

As noted earlier, rather different approaches may have performed equally well, for broad or 'generic' purposes, in larger-scale evaluations like DUC and NTCIR. But insofar as it is possible to distinguish better-performing systems from the rest, these appear to be ones that use more refined features, especially multi-word expressions, to characterise sentences. With more-demanding purposes, as in the query-oriented summarising in DUC 2005, the gains from greater interpretive sophistication appear to be larger, particularly since question analysis as well as source document processing is required.

Enriched statistical approaches: structures

In relation to the larger range of strategy options, two recent developments within the extractive approach are particularly important. The first is a more comprehensive use of source structure.

Thus systems may use sentence characterisations not merely to identify units and/or features, but for source representations in which structure is expressed and handed on for further processing, even though the final summary is wholly or at least primarily extractive. For example, parse trees that mark nominal structures in source sentences may be used not just as guides to source sentence scoring, but carried forward to guide text component selection for the output summary, as in Newsblaster (*NWBL*, McKeown et al. 2002).

But source structure here is still sentence-level structure. There is no ref-

erence to *discourse* structure beyond the statistical model of concept salience that lexical frequency or co-frequency supply. However there are richer, but still statistical, approaches to discourse structure than this minimal weak one. Thus Erkan and Radev (2004) use graph structures to determine sentence centrality, though these are only based on lexical relations between sentences. Others have used sentence structure as well, notably interpretations into logical forms, so lexical links between sentences are based on relations between logical form elements. Tucker and Sparck Jones (2005) use several network properties to identify sentences to select for the summary, and Vanderwende et al. (2004) and Lescovic et al. (2005) also use graphs over logical forms.

Statistical approaches to discourse structure can also be used to determine topic flow, not just topics and their connections, and thus order output. Boguraev and Kennedy (1999) segment source documents using lexical overlaps, and summarise segment by segment, showing pertinent extracts. Nakao (2000) uses lexical segmentation for hierarchical book summarisation.

Other moves to identify and use semantic/pragmatic discourse structure have exploited symbolically-defined structures, i.e. ones treating meaning explicitly rather than, as in the statistical case, implicitly. These have included local anaphoric structure based on distinctions like given/new. Thus Boguraev and Kennedy resolve anaphors so as to improve counting information for source units in interpretation. But attempts have also been made to use richer, and global, symbolic discourse structures. Most work has been done with Rhetorical Structure Theory. Miike et al. (1994) and Marcu (1999a, 2000) build RST source text trees by exploiting discourse marker expressions in particular, and use them to identify nucleus source clauses to extract for summaries, Miike et al. by weighting relations differentially, Marcu by scoring tree node dominance status. PALSUMM (Polanyi et al. 2004, Thione et al. 2004) build more abstract discourse trees using relations like subordination, and prune them to obtain summary text units. Teufel and Moens (1998, 2002) apply rhetorical, and specifically argument, categories to identify important source sentences.

In both such statistical and symbolic approaches source and summary representations are usually of the same general type, with the latter some selection drawn from the former, with varying transformational effort. In the system as a whole, most effort goes into interpretation.

But as emphasised in earlier summarising work (Endres-Niggemeyer et al. 1995); Sparck Jones 1995), there are different generic types of discourse structure, and many variants of each: very broadly, linguistic, world, and communicative types, each with top-down or bottom-up forms. RST and the PALSUMM model are both linguistic models of a general kind, but very different. Teufel and Moens' categories are also linguistic but broadly genre oriented to technical papers. Marcu (1998) suggested that evaluation had not shown that these richer symbolic discourse structures were of real use, especially as they cannot be identified very reliably. But work with them overall has been very limited, and in many cases has not reached evaluation stage, e.g. for Carberry et al. (2004)'s use of structures defined by communicative intentions.

However more specific application-oriented structures may be more effective,

especially within particular domains, for example for legal sources (Grover et al. 2003; Farzinder and Lapalme 2004), and where summarising verges on classical information extraction. With applications when the type of material to be extracted is pre-specified, much of the source can normally be ignored, there may be no requirement for a cohesive or coherent summary text, and structure clues may be clearer because domain-, i.e. world-, related. Source interpretation is designed to identify the material to fill template slots, which may be less (Farzinder and Lapalme) or more fine-grained (McKeown et al. 1998, Elhadad and McKeown 2001). Such application cases further illustrate different discourse structure types. Thus Moens and Dumortier (2000)'s text grammar is a linguistic structure; McKeown et al. and Elhadad and McKeown's main structure is a medical world one; Zhou and Hovy (2005) illustrate an input/response communicative structure.

The work with symbolic structures also illustrates top-down model forms, as in McKeown at al.'s schemas, others bottom-up ones, as in RST; and individual systems may combine several types of structure: thus McKeown et al. and Elhadad and McKeown use linguistic structure features of the source, associated with the domain, to identify source material for the domain-based source representation.

These richer structures may be used only for interpretation and exploited in transformation to select the material for the summary, leading (logically) to summary representations which simply identify the extracts to deliver. Marcu and PALSUMM use their linguistic structure for this, White and Cardie (2002) group and feed information from their event templates into sentence selection. However the (types of) structure used for source interpretation may also be used to organise the output summary. Thus McKeown et al. use domain structure to order blocks of output, and linguistic structure to order individual sentences, and Lapata and Barzilay (2005) use two types of linguistic structure. In some cases source structure like a template may be carried forward, perhaps with some heading relabelling, along with slot-filling text, for the output summary (Farzinder and Lapalme). In other cases, especially in multi-document summarising, source material is not just copied but reformulated and reordered (Elhadad and McKeown).

Harabagiu and Lacatusu (2005) illustrate both the possibilities and complexities of working with discourse structure. But they also show the value (for multi-document summarising) from working with large-scale and even general-purpose models, and from using explicit symbolic structure as well as implicit statistical structure. They use statistical topics along with syntactic and semantic patterns to identify themes expressing conceptual (i.e. world-referring) content. Themes can combined using linguistic and content relations, and represented as graphs which can be exploited to identify key source content and order it for output.

Comments on extractive summarising

Many variations of the extractive approach have been tried in the last ten

years. However it is hard to say how much greater interpretive sophistication, at sentence or text level, contributes to performance. Multi-party evaluations have generally not been challenging or discriminating enough to determine this. Single system evaluations for specific applications do not always offer comparison links, and many designs or implementations have not been evaluated. The assessment of added value is complicated by the fact that systems are described more in terms of local parameter choices than by underlying model properties.

It is true that the growth of research has stimulated the use of $tf * idf$-style benchmarks. This provides some, but weak, comparative information, and may be more useful in encouraging a focus on application specifics than in guiding larger choices. Conscious comparisons between types of approach, as in Harabagiu and Lacatusu (2005) are thus especially instructive. Some may argue that concentrating on what a summarising system actually does is all that is necessary for assessment and development. But against this, a more careful model-based analysis should help to understand individual process roles, and to relate these to individual task conditions.

Machine learning

This applies even though the second major development in extractive summarising, the use of machine learning, seems to imply that bootstrapping from data can deliver effective systems without the effort of model analysis.

Given the range of possible features for source characterisation, it is natural to ask whether machine learning can choose appropriate features, feature weights and feature combinations, Kupiec et al. (1995) and Teufel and Moens (1997, 2002) illustrate straightforward applications of this idea. But richer source characterisations can also be used, as in Marcu's use of RST-parsed source information to guide extraction, and in Leskovec et al. (2005)'s use of analysed sentence triples and graph relations to train an SVM classifier. Marcu and Echihabi (2002) could identify some discourse relations even with unsupervised learning. Machine learning can also be used for output, as in Barzilay and Lapata (2005)'s training for output sentence ordering.

In these cases machine learning has a preliminary and support role, to identify the information to be applied at specific process stages, e.g. in interpretation or generation. It may also be used in hybrid systems to supply resources or motivate particular processes.

But machine learning can be pushed further as a fundamentally statistical process that seems to conflate the three-stage model. Thus Banko et al. (2000) applied Language Modelling to training source-summary pairs to identify correlations that could be applied to new sources both to select output summary ngrams and to determine their ordering, in a drastic source text to summary sentence compression operation. Berger and Mittal (2000a) showed that FAQ data could be used as proxy training data where regular data is not available.

In this strategy, there seems to be no need to address summary fitness for purpose explicitly since it follows the data, or to be more than minimal about what might be a feature. But this presupposes that the training summaries

are fit for purpose, which may not be known, and less minimal features may be more effective. Thus Knight and Marcu (2002) uses symbolic parse data for sentence compression, and Daumé and Marcu (2002) compress whole texts using sentence syntax and RST information, so hybrid statistical-symbolic summarising replaces the basic statistical string-based one. It is certainly not clear that in practice the most basic approach adopted in Banko et al., which was applied to produce single 'phrase/sentence' summaries, could deliver coherent longer text. Moreover, even with the most minimal approach, there are logically the model's interpretation and generation stages, taking in or producing character/word strings, though with the major work shifted to selecting and ordering them in transformation, and in the hybrid cases there is more interpretive work.

The compression methods just noted are exciting as technology, particularly since they appear to satisfy the generic requirement for summarising as a condensation process. They are attractive, that is, because they seem to capture what is needed without any explicit, or at any rate in-depth, characterisation of source and summary meaning properties and their relationships. But what has been done so far is very limited in relation to the potential range of summary task conditions.

(I distinguish statistical compression, as discussed here, from *compaction*, where e.g. unimportant words, or syntactic substructures, are deleted from extracted sentences. Compaction is a valuable element in extractive summarising since it typically improves both content focus and expressive coherence, and it figures in more systems than those mentioned as doing pruning.)

Non-extractive strategies

In contrast to these extractive approaches, the second group jettisons the idea that summarising is about selectively reproducing some of the source text (beyond individual lexical items). Even approaches that may prune and merge source sentences or constituents (e.g. Newsblaster, McKeown et al. 2002), are essentially extractive. Non-extractive summaries are normally referred to as *abstracts* but this is too narrow a term, as there are other forms of non-extractive summary, like synopses or reviews.

The non-extractive methods that have been investigated have (as early exemplified by DeJong 1982) generally sought to dig well below the source linguistic surface to identify important conceptual content. They have as a corollary engaged with deeper sentence analysis and with overall discourse analysis so, for example sentence analysis to logical forms is deployed to find discourse relations like 'Consequence' that play a part in signalling concept status and significance. As noted for extractive summarising using discourse structure, such interpretive approaches, especially when highly selective for task reasons, may also be used for extracted output. But in general, the task requirements for non-extractive summaries can be expected to require novel output text generation.

As with extractive summarising, the different model types discussed in Sparck Jones (1995) can be used for non-extractive summarising, for example world

structure in DeJong and communicative structure in Reithinger et al. (2000). Linguistic structure does not seem to have been thus used, but linguistic structure based on logical forms, as in Tucker and Sparck Jones (2005), could lead to wholly new text. Both top-down and bottom-up approaches have been deployed. Reithinger et al.'s negotiation objects are simple top-down communicative structures, Hahn and Reimer (1999)'s domain relations are more bottom-up than top-down. Hahn and Reimer do not deliver text output, but clearly see it as one way of exploiting their internal information summaries, and doing so would naturally produce novel text. Again, as with the extractive cases, different structure types may be combined, e.g. Hahn and Reimer use both domain relations and statistical text ones. Saggion and Lapalme (2002) exploit a mix of domain-oriented genre concepts and relations and communicative ones (for indicating or informing), using templates and pattern matching to identify key source content. Instantiated templates as the source representation are selectively transformed for a summary representation as a standard genre-oriented presentational schema from which formatted output is produced. All of this work fits the three-stage processing model well, with elaborate deep source representations largely substituting for source texts, and also deep summary ones. However if the source models are also intrinsically selective, as in DeJong, transformation may be minimal. Tucker and Sparck Jones, Hahn and Riemer, and Saggion and Lapalme illustrate richer transformations.

There has been relatively little non-extractive summarising in the last decade, so it is harder to draw any conclusions about what it shows, or to compare it for task pertinence and performance with non-extractive approaches. This is not surprising, because non-extractive strategies are more effort, may not be readily portable, and are problematic for wide ranging source material like news. However as Hahn and Reimer point out, deeper source representations may lend themselves to a wider range of task applications for given inputs.

Comments on system characteristics

It is evident that, whether extractive or non-extractive, the systems developed in the last decade vary widely in complexity, detail and distribution of processing effort. They differ, in particular, in the treatment of discourse structure as a key guide in summarising, even if many extractive systems make use of some kind of lexically-based salient topic identification. These differences apply to the generic types of structure used, their sophistication, and to whether the system uses one or more types. Thus where some systems have just one type of statistical salience structure, as embodied in $tf*idf$-type weighting, Elhadad and McKeown (2001), for example, use structures of two types: source text linguistic structure to identify pertinent material and a domain structure to represent this, with a different derived domain structure for the summary representation and a further linguistic structure to organise text output.

These differences are sometimes attributable to specific task and context requirements (Elhadad and McKeown), sometimes to a desire to raise summarising

standards in general (Harabagiu and Lacatusu 2002). But in general, the scatter of systems, when taken together with the potential range of task applications, makes it impossible now to draw any very concrete, comparative conclusions about real versus trivial differences between approaches, about whether strategies fit tasks, or about the contribution that discourse structure analysis and representation make. In some cases choices of strategy and structure are clearly motivated by the task, e.g. in Farzinder and Lapalme (2004)'s legal application. But the lack of comparative evaluation, and even any evaluation, makes it difficult to judge strategies' relative merits. Indeed even where there has been quite careful evaluation, as in DUC, it does not support strong conclusions about system strategies. It appears that more intensive use of richer structures may be useful. But this can only be a very tentative conclusion, in part based on the fact that discourse has structure so it must be helpful for NLIP tasks. It may equally be that individual applications differ so much in their factor detail that we cannot expect much strategy portability, and have to settle for weaker generalisation.

Factor influences on strategy choices

Thus in reviewing recent work, while some approaches to summarising follow directly from the task context, for example McKeown et al. (1998), much more seems to follow either from some generic view of summarising without any significant task context analysis, or from current fashion, or from available or feasible technology. This last, in particular, encourages ad hoc extractive approaches that may suffice in practice, as with simple indicative summaries for document assessment in retrieval.

But as this implies, evaluation so far has not in general forced any thoroughgoing factor analyses, especially to determine how fine-grained strategy choices have to be. For example do different sources require different nominal group treatments in interpretation and transformation, or individual author styles require specific responses? The limited evidence so far suggests that in some applications this may be the case, so there is much more work to do.

Exemplar systems

I have so far considered broad classes of system, and used individual systems to illustrate particular points. This section offers a complementary view through brief accounts of exemplar systems as wholes. I have limited it to systems without the leverage of query orientation, or with very reductive headline output, and to ones that have been subject to some robustness testing, e.g. in DUC. The systems are well known but are useful in emphasising the wide variety of approaches adopted, even though performance levels are less varied. Most are for multi-document summarising because the field has focused on this, but single-document summarising is also covered.

MEAD (Radev et al. 2001b, 2004)

MEAD is an essentially statistical system for either single- or multi-document summarising. For single documents or (given) clusters it computes centroid topic characterisations using $tf * idf$-type data. It ranks candidate summary sentences by combining sentence scores against centroid, text position value, and $tf*idf$ title/lead overlap. Sentence selection is constrained by a summary length threshold, and redundant new sentences avoided by checking cosine similarity against prior ones.

The forms of representation are simple term vectors and score sets, and processing is equally simple vector comparison and score computation. But MEAD performed respectably in DUC 2001 and 2002, is used as a component of other systems, e.g. NewsInEssence (*NIE*), and is a public domain system.

Newsblaster (NWBL, McKeown et al. 2002)

Newsblaster is primarily statistical, but with significant symbolic elements. As a fully operational public system, it includes operations and addresses concerns that do not figure in the mostly experimental summarising literature. Thus it includes initial processes to identify news stories and cluster them, and output processes to add e.g. images and links for users to sources and other resources. Clustering is multi-level, using both $tf * idf$-type and syntactic features, with top-level assignment to broad news categories. Stories are typed as event, person, or 'other'.

For events, for example, a document cluster is processed to find similar paragraphs, defining themes, using $tf * idf$ weighting and symbolic features. Theme sentences are then symbolically analysed and their parses compared to identify syntactically and semantically similar components. Similar parse trees are fused and, because the content repeats they embody indicate importance, taken for the summary. The selected parse constituents are ordered by original appearance and processed by a text generator, combining and filling them out for complete output sentences. Summaries for the other types use a generic content model to guide summarising, within the same framework.

This is a sophisticated system, which includes training for features to use, and illustrates the range of representations and processes that may be used in summarising. The source representation is the set of theme paragraphs with their sentence parses. The major work is in transformation, comparing, fusing and ordering parse parse tree constituents, with further work in generation. Both representations are symbolic, with the summary one remoter and deeper than the source one. The core system performed well in DUC 2001 and 2002, and the system as a whole in a specific evaluation (McKeown et al. 2005).

GISTexter (Harabagiu and Lacatusu 2002)

GISTexter is also a sophisticated system, but with a very different basis. It is designed to produce both single- and multi-document summaries, and both extracts and abstracts. The single-document summarising is simple, the multi-document of real interest. The abstracts are in fact extracts, labelled as ab-

stracts because produced with information extraction (IE) techniques.

Thus GISTexter multi-document summarising uses IE-style templates, either from a prior set or by ad hoc generation. Set document sentences are parsed and co-references resolved, supplying the information to map source snippets along with their co-reference relations into template slots using pattern rules. The templates for a document set are classified, using the reference data, as about main or subsidiary events. The summary generation stage, guided by template class, references, the amount of filler material, and desired length, invokes source sentences for the template snippets. These are output, perhaps pruned for length, in source order and with suitable reference forms.

The template generator is of interest as a response to open-domain sources like news. It exploits WordNet to identify topic relations that define semantic roles for key source lexical items. These bootstrapped linguistic templates are less powerful than hand-crafted ones, but still useful.

GISTexter performed well in DUC 2002. It is an elaborate, resource-rich system, involving complex parsing, reference resolution, and template manipulation, but also a flexible one. The source representation is given by the filled templates. Transformation operates over these to select key source content indicators, and generation delivers appropriately tweaked corresponding source sentences. GISTexter was replaced by Lite-GISTexter in DUC 2003, but this was because the task changed (Lacatusu et al. 2003).

Verbmobil (Reithinger et al. 2000)

Reithinger et al. illustrate a very different summarising situation and approach, for multi-lingual spoken dialogues in a limited (travel) domain. Summarisation is one function in a system primarily devoted to speech transcription and translation.

Transcribed utterances are parsed to extract dialogue acts and their domain content. Content is mapped into domain content templates, with dialogue act operators, and these units are grouped into 'negotiation objects', e.g. PROPOS[AL], which are refined as the dialogue progresses. Summaries are user requested, and based on the most complete negotiation object for each major travel category (accommodation, travelling). They are generated in e.g. German or English, using discourse and sentence planning, with category content packaged with suitable verb fillers and discourse control to ensure correct focusing. Reithinger et al. report a very limited evaluation, though Verbmobil as a whole was also evaluated, for translation.

Reithinger et al.'s summarising illustrates a very different context from the previous ones, geared to dialogue and helped by a well-specified domain. It is based on rich symbolic processing, with some statistical help to identify dialogue acts, and exploits both a domain world model in its templates and communicative dialogue and negotiation models. The source representation is a set of negotiation objects, the summary representation a selected subset of these. The main processing is in input interpretation, with some transformation and a little more in generation.

6 Conclusion

The status, and state, of automatic summarising has radically changed in the last ten years. There is a large research community, and there are operational systems working with open-domain sources in varied conditions. Some systems are simple, e.g. Web search engine ones, but useful in retrieval; others are sophisticated, e.g. Newsblaster, but their uses, users and value are essentially unknown.

Summarising has benefited from work on neighbouring tasks, notably retrieval and question answering. It has benefited from training and test data. Most importantly, it has benefited from the evaluation programmes of the last ten years. These have been significant both for the system work they have stimulated and the results obtained, and for the development of evaluation methodologies and a growing awareness of the need for proper task specification and performance assessment.

In relation to summarising techniques themselves, this wave of work has been useful in exploring the possibilities and potential utilities of extractive summarising, and specifically statistical and/or shallow symbolic methods that do not require heavy model instantiation, for example in domain ontologies. There is some evidence such techniques can deliver useful goods where the summary requirements are modest, and hybrid techniques a little more than purely statistical ones.

There is no reason, therefore, to suppose that summarising research and development will not continue.

However, against this, the work and evaluations done so far have been limited and miscellaneous when compared with the views of the summarising space discussed at the Dagstuhl Seminar in 1993 (Endres-Niggemeyer et al. 1995). The work on extractive summarising has picked the low-hanging fruit, and the overall trend has been more technological than fundamental. There has been little work on the deep approaches that appear to be needed if source-to-summary condensation requires radical transformation of content and expression. This is not surprising: as Marcu (1999, 2000)'s experiments suggest, we do not know, except with heavy application-specific guidance, how to automate such processes.

As a result, given the difficulty of specifying tasks, capturing their pertinent factor conditions, and evaluating system performance for task, we cannot say much about the types of summarising strategy, with their forms of represntation and process, that suit tasks. There is a lesson here from TREC (Voorhees and Harman 2005), which began with a conventional view of retrieval but branched out, pushing existing technology and developing new. DUC and its sister programmes are beginning this for summarising, but have much further to go. To make progress on the road we must drive future research through more challenging formulations of the task, because these will call for more analysis of and response to the factors that affect summarising, and more careful evaluation distinctions than the simple intrinsic/extrinsic one.

References

Workshops (in temporal order):
ACL-97: *Intelligent scalable text summarisation*, (Ed. I. Mani and M. Maybury), ACL, 1997.
AAAI-98: *Intelligent text summarisation*, (Ed. E. Hovy and D. Radev), AAAI Spring Symposium, AAAI, 1998.
ANLP/NAACL-00: *Automatic summarisation*, (Ed. U. Hahn, C.-Y. Lin and D. Radev), ACL, 2000.
NAACL-01: *Automatic summarisation*, (Ed. J. Goldstein and C.-Y. Lin), ACL, 2001.
ACL-02: Text summarisation, (Ed. U. Hahn and D. Harman), ACL, 2002.
ACL-03: Multilingual summarisation and question answering, ACL, 2003.
HLT-NAACL-03: *Text summarisation*, (Ed. D. Radev and S. Teufel), ACL, 2003.
ACL-04: *Text summarisation branches out*, (Ed. M.-F. Moens and S. Szpakowicz), ACL, 2004.
ACL-05: *Intrinsic and Extrinsic Evaluation Measures for MT and/or summarisation* (Ed. J. Goldstein et al.), ACL, 2005.

Amigo, E. et al. (2005) 'QARLA: a framework for the evaluation of text summarisation systems', *ACL 2002: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, 2005, 280-289.

Ando, R.K. et al. (2000) 'Multi-document summarisation by visualising topical content', *ANLP/NAACL-00*, 2000, 79-88.

Aone, C. et al. (1997) 'A scalable summarisation system using robust NLP', *ACL-97*, 1997, 66-73.

Banko, M. Mittal, V. and Witbrock, M. (2000) 'Headline generation based on statistical translation', *ACL 2000: Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, 2000, 318-325.

Barzilay, R. and Elhadad, M. (1999) 'Using lexical chains for text summarisation', in Mani and Maybury (1999), 110-121.

Barzilay, R. and Lapata, M. (2005) 'Modelling local coherence: an entity-based approach', *ACL 2000: Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, 2000, 318-325.

Berger, J. and Mittal, V. (2000) 'Query-relevant summarisation using FAQs', *ACL 2000: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, 2005, 141-148. (2000a)

Berger, J. and Mittal, V. (2000) 'OCELOT: a system for summarising web pages', *Proceedings of the 23rd Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2000)*, 2000, 144-151. (2000b)

BMJ: *British Medical Journal*,
http://www.bmjjournals.com (visited April 2006).

Boguraev B. et al. (1998) 'Dynamic presentation of document content for rapid on-line skimming', *AAAI-98*, 1998, 111-117.

Boguraev, B. and Kennedy, C. (1999) 'Salience-based content characterisation of text documents', in Mani and Maybury (1999), 99-110.

Boguraev, B., Bellamy, R. and Swart, C. (2001) 'Summarisation miniaturisation: delivery of news to hand-helds', *NAACL-01*, 2001, 99-108.

Brandow, R., Mitze, K. and Rau, L.F. (1995) 'Automatic condensation of electronic publications by sentence selection', *Information Processing and management*, 31 (5), 1995, 675-686. Reprinted in Mani and Maybury (1999).

Carberry, S. et al. (2004) Extending document summarisation to information graphics', *ACL-04*, 2004, 3-9.

Carbonell, J. and Goldstein, J. (1998) 'The use of MMR and diversity-based reranking for reordering documents and producing summaries', *Proceedings of the 21st Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1998)*, 1998, 335-36.

Chakrabarti, S., Joshi, M. and Tawde, V. (2001) 'Enhanced topic distillation using text, markup tags, and hyperlinks', *Proceedings of the 24th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001)*, 2001, 208-216.

Chan, S.W.K. et al. (2000) 'Mining discourse markers for Chinese text summarisation', *ANLP/NAACL-00*, 2000, 11-20.

Christel, M.G. et al. (2002) 'Collages as dynamic summaries for news video', *Proceedings of ACM Multimedia 2002*, 2002.

Corston-Oliver, S. (2001) 'Text compaction for display on very small screens', *NAACL-2001*, 2001, 89-98.

Corston-Oliver, S. et al. (2004) 'Task-focused summarisation of email', *ACL-04*, 2004, 43-50.

Damianos, L. et al. (2002) 'MiTAP for bio-security: a case study', *AI Magazine*. 23 (4), 2002, 13-29.

Daumé, H. and Marcu, D. (2002) 'Noisy channel model for document compression', *ACL 2002: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002, 449-456.

Daumé, H. and Marcu, D. (2004) 'Generic sentence fusion is an ill-defined task', *ACL-04*, 2004, 96-103.

DeJong, G. (1982) 'An overview of the FRUMP system', in *Strategies for natural language processing*, (Ed. W.G. Lehnert and M.D. Ringle), Hillsdale, NJ: Lawrence Erlbaum), 1982, 149-176.

Dorr, B., Zajic, D. and Schwartz, R. (2003) 'Hedge Trimmer: a parse-and-trim approach to headline generation', *HLT-NAACL-03*, 2003, 1-8.

Dorr, B.J. et al. (2005) 'A methodology for extrinsic evaluation of text summarisation', *ACL-05*, 2005, 1-8.

Douzidia, F.S. and Lapalme, G. (2004) 'Lakhas, an Arabic summarising system', *DUC 2004*, 2004, 128-135.

DUC: Proceedings of the DUC Workshops 2001-2005,
http://duc.nist.gov/ (visited April 2006).

Earl, L.L. (1970) 'Experiments in automatic indexing and extracting', *Information Storage and Retrieval*, 6, 1970, 313-334.

Edmundson, H.P. (1969) 'New methods in automatic extracting', *Journal of the ACM*, 16 (2), 1969, 264-285. Reprinted in Mani and Maybury (1999).

Elhadad, N. and McKeown. K.R. (2001) 'Towards generating patient specific summaries of medical articles', *NAACL-01*, 2001, 32-40.

Endres-Niggemeyer, B., Hobbs, J. and Sparck Jones, K. (Eds.) (1995) *Summarising text for intelligent communication*, Dagstuhl-Seminar-Report; 79 (Full version), IBFI GmbH Schloss Dagstuhl, Germany, 1995.

Endres-Niggemeyer, B. (1998) *Summarising information*, Berlin: Springer, 1998.

Erkan, G. and Radev, D. (2004) 'LexRank: graph-based centrality as salience in text summarisation', *Journal of Artificial Intelligence Research*, 22, 2004, 457-479.

Evans, R. et al. (1995) 'POETIC: a system for gathering and disseminating traffic information', *Journal of Natural Language Engineering*, 1 (4), 1995, 363-387.

Farzinder, A. and Lapalme, G. (2004) 'Legal text summarisation by exploration of the thematic structure and argumentative roles', *ACL-04*, 2004, 27-34.

Farzinder, A. and Lapalme, G. (2005) 'Production automatique du résumé de textes juridiques: évaluation de qualité et d'acceptabilité', *TALN 2005*, Dourdan, France, 2005, Vol. 1, 183-192.

Filatova, E. and Hatzivassiloglou, V. (2004) 'Event-based extractive summarisation', *ACL-04*, 2004, 104-111.

Furui, S. (2005) 'Spontaneous speech recognition and summarisation', *Proceedings of the Second Baltic Conference on Human Language Technologies*, Talinn, 2005, 39-50.

Futrelle, R. (2004) 'Handling figures in document summarisation', *ACL-04*, 2004, 61-65.

Grefenstette, G. (1998) 'Producing intelligent telegraphic text reduction to provide an audio scanning service for the blind', *AAAI-98*, 1998, 111-117.

Grewal, A. et al. (2003) 'Multi-document summarisation using off-the-shelf compression software', *HLT-NAACL-03*, 2003, 17-24.

Grover, C., Hachey, B. and Korycinski, C. (2003) 'Summarising legal texts: sentential tense and argumentative rules', *HLT-NAACL-03*, 2003, 33-40.

Hahn, U. (1990) 'Topic parsing: accounting for text macro structures in full-text analysis', *Information Processing and Management*, 26, 1990, 135-170.

Hahn, U. and Reimer, U. (1999) 'Knowledge-based text summarisation: salience and generalisation for knowledge base abstraction', in Mani and Maybury (2000), 215-222.

van Halteren, H, and Teufel, S. (2003) 'Examining the consensus between human summaries: initial experiments with factoid analyses', *HLT-NAACL-03*, 2003, 57-64.

Hand, T.F. 'A proposal for task-based evaluation of text summarisation systems', *ACL-97*, 1997, 31-38.

Harabagiu, S. and Lacatusu, F. (2002) 'Generating single and multi-document summaries with GISTexter' *DUC 2002*. 2002, 30-38.

Harabagiu, S. and Lacatusu, F. (2005) 'Topic themes for multi-document summarisation', *Proceedings of the 28th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005)*, 2005, 202-209.

Harman, D. and Over, P. (2004) The effects of human variation in DUC summarisation evaluation', *ACL-04*, 2004, 10-17.

Hirao, T., Sasaki, Y. and Isozaki, H. (2001) 'An extrinsic evaluation for question-biased text summarisation on QA tasks', *NAACL-01*, 2001, 61-68.

Hori, C., Hirao, T. and Isozaki, H. (2004) 'Evaluation measures considering sentence concatenation for automatic summarisation by sentence or word extraction', *ACL-04*, 2004, 82-88.

Hovy, E. and Lin, C.-Y. (1999) 'Automated text summarisation in SUMMARIST', in Mani and Maybury 2000, 81-94.

IPM 1995): Sparck Jones, K. and Endres-Niggemeyer, B. (Eds.) (1995) 'Summarising text', Special Issue, *Information Processing and Management*, 31 (5), 1995, 625-784.

Jing, H. et al. (1998) 'Summarisation evaluation methods: experiments and analysis', *AAAI-98*, 1998, 60-68.

Jing, H. (2002) 'Using hidden Markov modelling to decompose human-written summaries', *Computational Linguistics*, 28 (4), 427-443.

Jordan, D. et al. (2004) An evaluation of automatically generated briefings of patient status', *MEDINFO 2004*, (Ed. M. Fieschi et al.), Amsterdam: IOS Press, 2004, 227-231.

Knight, K. and Marcu, D. (2002) 'Summarisation beyond sentence extraction: a probabilistic approach to sentence compression', *Artificial Intelligence*, 139, 2002, 91-107,

Kolluru, B. and Gotoh, Y. 'On the subjectivity of human authored short summaries', *ACL-05*, 2005.

Kupiec, J. Pedersen, J. and Chen, F. (1995) 'A trainable document summariser', *Proceedings of the 18th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1995)*, 1995, 68-73.

Lacatusu, V.F., Parker, P. and Harabagiu, S.M. (2003) 'Lite-GISTexter: generating short summaries with minimal resources', in *DUC 2003*, 122-128.

Lacatusu, F. et al. (2005) 'Lite-GISTexter at DUC 2005', *DUC 2005*, 2005, 88-94.

Lam-Adelsina, A. and Jones, G.F.J. (2001) 'Applying summarisation techniques for term selection in relevance feedback', *Proceedings of the 24th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001)*, 2001, 1-9.

Lapata, M. and Barzilay, R. (2005) 'Automatic evaluation of text coherence: models and representations', *Proceedings of IJCAI*, 2005.

Leskovec, J., Milic-Frayling, N. and Grobelnik, M. (2005) 'Impact of linguistic analysis on the semantic graph coverage and learning of document extracts',

Proceedings of the AAAI, 2005.

Lin, C.-Y. and Hovy E. (2000) 'The automated acquisition of topic signatures for text summarisation', *Proceedings of 18th International Conference on Computational Linguistics (COLING 2000)*, 2000, 495-501.

Lin, C.-Y. and Hovy, E. (2002) 'Automated multi-document summarisation in NeATS', *Proceedings of the Human Language Technology Conference (HLT 2002)*, 2002, 50-53. (2002a)

Lin, C.-Y. and Hovy, E. (2002) 'Manual and automatic evaluation of summaries', *ACL-02*, 2002, 45-51. (2002b)

Lin, C.-Y. and Hovy, E. (2003) 'The potential and limitations of automatic sentence extraction for summarisation', *HLT-NAACL-03*, 2003, 73-80.

Lin, C.-Y. (2004) 'ROUGE: a package for automatic evaluation of summaries', *ACL-04*, 2004, 74-81.

Luhn, H.P. (1958) 'The automatic creation of literature abstracts', *IBM Journal of Research and Development*, 2 (2), 1964, 159-165. Reprinted in Mani and Maybury (1999).

Mani, I. and Bloedorn, (1997) 'Multi-document summarisation by graph search and matching', *Proceedings of the Annual Conference of the AAAI*, 1997, 622-628.

Mani, I. and Maybury, M.T. (Eds.) (1999) *Advances in automatic text summarisation*, Cambridge MA: MIT Press, 1999.

Mani, I., Concepcion, K. and van Guilder, G. (2000) 'Using summarisation for automatic briefing generation',rr *ANLP/NAACL-00*, 2000, 89-98.

Mani, I. (2001) *Automatic summarisation*, Amsterdam: John Benjamins, 2001.

Mani, I. et al. (2002) 'SUMMAC: a text summarisation evaluation', *Natural Language Engineering*, 8 (1), 2002, 43-68.

Marcu, D. (1998) 'To build text summaries of high quality, nuclearity is not sufficient', *AAAI-98*, 1998, 1-8.

Marcu, D. (1999) 'Discourse trees are good indicators of importance in text', in Mani and Maybury (1999), 123-136. (1999a)

Marcu, D. (2000) *The theory and practice of discourse parsing and summarisation*, Cambridge MA: MIT Press, 2000.

Marcu, D. and Gerber, L. (2001) 'An inquiry into the nature of multidocument abstracts, extracts and their evaluation', *NAACL-01*, 2001, 2-11.

Marcu, D. and Echihabi, A. (2002) 'An unsupervised approach to recognising discourse relations', *ACL 2000: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002, 368-375.

Maybury, T. (1995) 'Generating summaries from event data', *Information Processing and Management*, 31 (5), 1995, 736-751. Reprinted in Mani and Maybury (1999).

Maynard, D. et al. (2002) 'Using a text engineering framework to build and extendable and portable IE-based summarisation system', *ACL-02*, 2002, 19-26.

McKeown, K., Robin, J. and Kukich, K. (1995) 'Generating concise natural language summaries', *Information Processing and Management*, 31 (5), 1995, 703-733. Reprinted in Mani and Maybury (1999).

McKeown, K., Jordan, D. and Hatzivassiloglou, V. (1998) 'Generating patient-specific summaries of online literature', *AAAI-98*, 1998, 34-43.

McKeown, K. et al. (2001) 'Columbia multi-document summarisation: approach and evaluation', *DUC 2001*, 2001.

McKeown, K.R. et al. (2002) 'Tracking and summarising news on a daily basis with Columbia's Newsblaster', *Proceedings of the Human Language Technology Conference (HLT 2002)*, 2002.

McKeown, K. et al. (2005) 'Do summaries help? A task-based evaluation of multi-document summarisation', *Proceedings of the 28th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005)*, 2005, 210-217.

Merlino, A. and Maybury, M. (1999) 'An empirical study of the optimal presentation of multimedia summaries of broadcast news', in Mani and Maybury (1999), 391-403.

Miike, S. et al. (1994) 'A full-text retrieval system with a dynamic abstract generation function', *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR 94)*, 1994, 152-161.

Minel, J.-L., Nugier, S. and Piat, G. (1997) 'How to appreciate the quality of automatic text summarisation', *ACL-97*, 1997, 25-30.

Moens, M.-F., Yttendaele, C. and Dumortier, J. (1997) 'Abstracting of legal cases: the SALOMON experience', *Proceedings of the Sixth International Conference on Artificial Intelligence and the Law*, ACM, 1997, 114-122.

Moens, M.-F. and Dumortier, J. (2000) 'Use of a text grammar for generating highlight abstracts of magazine articles', *Journal of Documentation*, 56, 2000, 520-539.

Morris, A.H., Kasper, G.M. and Adams, D.A. (1992) 'The effects and limitations of automated text condensing on reading comprehension performance', *Information Systems Research*, 3 (1), 1992, 17-35. Reprinted in Mani and Maybury (1999).

Murray, G., Renals, S. and Carletta, J. (2005) 'Extractive summarisation of meeting recordings', *ACL-05*, 2005.

Nakao, Y. (2000) 'An algorithm for one-page summarisation of a long text based on thematic hierarchy detection', *ACL 2000: Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, 2000, 302-309.

NWBL: Newsblaster,
http://newsblaster.cs.columbia.edu/ (visited August 2006)

NIE: NewsInEssence,
http://www.newsinessence.com/ (visited July 2006)

Nobata, C., Sekine,. S. and Isahara, H. (2003) 'Evaluation of features for sentence extraction on different types of corpora', *ACL-03*, 2003.

NTCIR:
http://research.nii.ac.jp/ntcir/index-en.html (visited April 2006).

Oka, M. and Ueda, Y. (2000) 'Evaluation of phrase-representation summarisation based on information retrieval task', *ANLP/NAACL-00*, 2000, 59-68.

Okurowski, M.E. et al. (2000) 'Text summariser in use: lessons learned from real world deployment and evaluation', *ANLP/NAACL-00*, 2000, 49-58.

Over, P. and Yen, J. (2004) 'Introduction to DUC 2004. Intrinsic evaluation of generic news text summarisation systems', *DUC 2004*, 2004, 1-21.

Papernick, N. and Hauptmann, A.G. (2005) 'Summarisation of broadcast news video through link analysis of named entities', *AAAI Workshop on Link Analysis*, 2005.

Passonneau, R.J. et al. (2005) 'Applying the Pyramid method in DUC 2005', *Document Understanding Workshop (DUC) 2005*, 25-32.

Polanyi, L. et al. (2004) 'A rule-based approach to discourse parsing', *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, ACL, 2004, 108-117.

Pollock, J.J. and Zamora, A. (1975) 'Automatic abstracting research at Chemical Abstracts Service', *Journal of Chemical Information and Computer Sciences*, 15 (4), 1975, 226-232. Reprinted in Mani and Maybury (1999).

Radev, D.R., Jing, H. and Budzikowska, M. (2000) 'Centroid-based summarisation of multiple documents: sentence extraction, utility-based evaluation, and user studies', *ANLP/NAACL-00*, 2000, 21-30.

Radev, D., Fan, W. and Zhang, Z. (2001) 'WebInEssence: a personalised web-based multi-document summarisation and recommendation system', *NAACL-01*, 2001, 79-88. (2001a)

Radev, D.R., Blair-Goldensohn, S. and Zhang, Z. (2001) 'Experiments in single and multi-document summarisation using MEAD', *DUC 2001*, 2001. (2001b)

Radev, D. et al. (2003) 'Evaluation challenges in large-scale document summarisation', *ACL 2003: Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, 2003, 375-382.

Radev, D. et al. (2004) 'MEAD - a platform for multilingual summarisation', *Proceedings of LREC 2004*, 2004.

Rath, G.J., Resnick, A. and Savage, T.R. (1961) 'The formation of abstracts by the selection of sentences', *American Documentation*, 12 (2), 1961, 139-143. Reprinted in Mani and Maybury (1999).

Reithinger, N. et al. (2000) Summarising multilingual spoken negotiation dialogues', *ACL 2000: Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, 2000, 310-317.

Rother, C. et al. (2006) 'AutoCollage', *SIGGRAPH '06: ACM Transactions on Graphics*, 2006.

ROUGE: http://haydn.isi.edu/ROUGE/ (visited April 2006).

Rowley, J. (1982) *Abstracting and indexing*, London: Bingley: 1982.

Saggion, H. and Lapalme, G. (2000) 'Concept identification and presentation in the context of technical text summarisation', *ANLP/NAACL-00*, 2000, 1-10.

Saggion, H. and Lapalme, G. (2002) 'Generating informative-indicative summaries with SumUM', *Computational Linguistics*, 28 (4), 2002, 497-526.

Sakai, T. and Sparck Jones, K. (2001) 'Generic summaries for indexing in information retrieval', *Proceedings of the 24th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001)*, 2001, 190-198.

Salton, G. et al. (1997) 'Automatic text structuring and summarisation', *Information Processing and Management*, 33 (2), 193-207. Reprinted in Mani and Maybury (1999).

Santos, E., Mohamed, A.A. and Zhao, Q. 'Automatic evaluation of summaries using document graphs', *ACL-04*, 2004, 66-73.

Sato, S. and Sato, M. (1998) 'Rewriting saves extracted summaries', *AAAI-98*, 1998, 85-92.

SEE: Over, P. and Yen, J. (2004) 'Introduction to DUC 2004. Intrinsic evaluation of generic news text summarisation systems', *DUC 2004*, 2004, 1-21.

Silber, H.G. and McCoy, K.F. (2002) 'Efficiently computed lexical chains as an intermediate representation for automatic text summarisation', *Computational Linguistics*, 28 (4), 2002, 487-496.

Sparck Jones, K. (1995) 'Discourse modelling for automatic summarising', in *Travaux du Cercle Linguistique de Prague* (Prague Linguistic Circle Papers), vol 1, 1995, 201-227.

Sparck Jones, K. and Galliers, J.R. (1996) *Evaluating natural language processing systems*, Berlin: Springer, 1996.

Sparck Jones, K. (1999) 'Automatic summarising: factors and directions', *Advances in automatic text summarisation*, (Ed. I. Mani and M.T. Maybury), Cambridge MA: MIT Press, 1999, 1-14.

Sparck Jones, K. (2001) 'Factorial summary evaluation', in *DUC 2001*, 2001.

Sparck Jones, K. (2007) *Automatic summarising: a review and discussion of the state of the art*, Technical Report 679, Computer Laboratory, University of Cambridge, 2007.

Strzalkowski, T. et al. (1999) 'A robust practical text summariser', in Mani and Maybury (1999), 237-154.

SUMMAC: *TIPSTER Text Summarisation Evaluation Conference (SUMMAC)*,
http://www-nlpir.nist.gov/related_projects/tipster_summac/

Sun, J.-T. et al. (2005) 'Web-page summarisation using click-through data', *Proceedings of the 28th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005)*, 2005, 194-201.

Teufel, S. and Moens, M. (1997) 'Sentence extraction as a classification task', *ACL-97*, 1997, 58-65.

Teufel, S. and Moens, M. (1998) 'Sentence extraction and rhetorical classification for flexible abstracts', *AAAI-98*, 1998, 16-25.

Teufel, S. (2001) 'Task-based evaluation of summary quality: describing relationships between scientific papers', *NAACL-01*, 2001, 12-21.

Teufel, S. and Moens, M. (2002) 'Summarising scientific articles: experiments with relevance and rhetorical status', *Computational Linguistics*, 28 (4), 2002, 409-445.

Thione, G.L. et al. (2004) 'Hybrid text summarisation: combining external relevance measures with structural analysis', *ACL-04*, 51-55.

Tombros, A., Sanderson, M. and Gray, P. (1998) 'Adequacy of query biased summaries in information retrieval', *AAAI-98*, 1998, 44-52.

Tucker, R.I. and Sparck Jones, K. (2005) *Between shallow and deep: an experiment in automatic summarising*, Technical Report 632, Computer Laboratory, University of Cambridge, 2005.

Vandegehinste, V. and Pan, Y. (2004) 'Sentence compression for automated subtitling: a hybrid approach', *ACL-04*, 2004, 89-95.

Vanderwende. L., Banko,M. and Menezes, A. (2004) 'Event-centric summary generation', *DUC 2004*, 2004, 76-81.

Voorhees, E.M. (2005) 'Question answering in TREC', in Voorhees and Harman 2005, 243-257. (2005a)

Voorhees, E.M. (2005) 'Overview of the TREC 2005 question answering track', *The Fourteenth Text REtrieval Conference, TREC 2005*, National Institute of Standards and Technology, 2005, 233-257. (2005b)

Voorhees, E.M. and Harman, D.K. (2005) *TREC: Experiment and evaluation in information retrieval*, Cambridge MA: MIT Press, 2005.

Wasson, M. (2002) 'Using summaries in document retrieval', *ACL-02*, 2002, 37-44.

White, M. and Cardie, C. (2002) 'Selecting sentences for multi-document summaries with randomised local search', *ACL-02*, 2002, 9-18.

Witbrock. M.J. and Mittal, V.O. (1999) 'Ultra-summarisation: a statistical approach to generating highly condensed non-extractive summaries', *Proceedings of the 22nd Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1999)*, 1999, 314-315.

Witten, I.H. et al. (2000) 'KEA: Practical automatic keyphrase extraction', Working Paper 00/5, Department of Computer Science, University of Waikato, 2000.

Yu, J. et al. (in press)) 'Choosing the content of textual summaries of large time-series data sets', *Natural Language Engineering*, in press.

Zechner, K. (2001) 'Automatic generation of concise summaries of spoken dialogues in unrestricted domains', *Proceedings of the 24th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2001)*, 2001, 199-207.

Zechner, K. (2002) 'Automatic summarisation of open-domain multiparty dialogues in diverse genres', *Computational Linguistics*, 28 (4), 2002, 447-485.

Zhang et al., H. Chen, Z. and Cai. Q. (2003) 'A study for document summarisation based on personal annotation', *HLT-NAACL-03*, 2003, 41-48.

Zhou, L. and Hovy, E. (2005) 'Digesting virtual 'geek' culture: the summarisation of technical internet relay chat', *ACL 2002: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, 2005, 298-305.

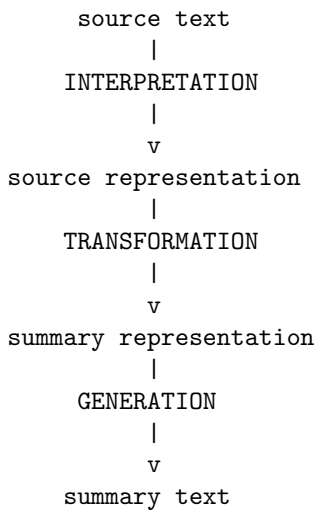

Figure 1: Schematic summary processing model for text

```
  input factors

     form - language, register, medium, structure, genre, length
     subject type
     unit
     author
     header (metadata)

[contrasted examples: archaeological paper, children's tale]

  purpose factors

     use
     audience
     envelope - time, location, formality, trigger, destination

[contrasted examples: emergency alert, literary review]

  output factors

     material - coverage, reduction, derivation, specialty
     style
     format - language, register, medium, structure, genre

[contrasted examples: bullet item list, prose paragraph]
```

Figure 2: Context factors affecting summarising

```
evaluation remit

establish :
  motivation - perspective, interest, consumer
  goal
  orientation, kind, type, form of yardstick, style, mode


evaluation design

identify :
  system (being evaluated) ends, context, constitution

determine :
  performance factors, ie environment variables, system parameters
  performance criteria, ie measures, methods

characterise :
  evaluation data

define :
  evaluation procedure
```

Figure 3: Decomposition framework for evaluation

```
intrinsic          semi-purpose
      ^               inspection eg for proper English
      |
      |            quasi-purpose
      |               comparison with models eg ngrams, nuggets
      |
      |            pseudo-purpose
      |               simulation of task contexts eg action scenarios
      |
      v            full-purpose
  extrinsic           operation in task context eg report writing
```

Figure 4: Evaluation relating to task context

```
Overall context: police reports of well-fed wombats sleeping on
roads and being a danger to traffic, prompting brief alerting summaries
to the local population through their newspaper.

Evaluation scenario sketch:
Remit :   Motivation -
            perspective - effectiveness (not cost)
            interest - system funders
            consumers - funders and builders
          Goal - brief warning alerts work
          Orientation - intrinsic for alerting setup
          Kind - investigation of response
          Type - black box
          Yardstick - police loudspeaker vans
          Style - indicative
          Mode - simple quantitative
Design :  Evaluation subject : alerting setup
          Subject's ends - avoid accidents
          Subject's context - geography, travel, accidents, wombats ...
          Subject's constitution - alerts, locals ...
          Performance factors :
            Environment variables -
              frequency of alerts, News sales, literacy of locals ...
            Setup parameters -
              summary features (eg length), alert repeats over pages ...
          Performance assessment :
            Criteria - success in alerting
            Measures - wombats avoided
            Methods - age, time etc breakdowns
          Evaluation data :
            data on alerts - number, topics, repeats ...
            data on locals - number, News exposures ...
            questionnaire responses
          Evaluation procedure :
            design and pilot questionnaire
            identify samples of locals
            set times for giving questionnaire
            log and score answers

Issues of detail (example) :
    population sampling;  questionnaire design
Evaluation variants : intrinsic - text beats graphics
                      extrinsic - saves police time on wombat accidents

Alternative purpose: factual summaries for research database on wombats
 Evaluation : Goal - establish summaries informative for researchers ...
              Design - determine database use for wombat papers ...
```


Figure 5: Examples of specific summarising contexts and evaluations

BMJ 2006; 332; 334-335

Objective
To describe the distribution of mortality among internally displaced persons ....

Design
Cross sectional household survey with retrospective cohort analysis of mortality.

Setting
Camps for internally displaced persons ....

Participants
3533 people from 859 households ....

Main outcome measures
All cause death and number of missing people.

Results
446 deaths and 11 missing people were reported after the 2004 tsunami, ...

Conclusions
Most mortality after the 2004 tsunami occurred within the first few days of the disaster and was low in the study area.

Figure 6: *BMJ* summary example

```
(Road Map 1)
DUC-01
     news material
     summaries - single documents, short
               - multiple documents, various lengths
     generic summaries (reflective, general-purpose)
  evaluation intrinsic :
     comparators - human summaries (reference)
                 - source openings (baseline)
     text quality (e.g. grammaticality)                          semi-purpose
     reference unit coverage (simple 'propositions')             quasi-purpose
  results : baselines <= systems < humans
            systems giving extracts, not junk, but not good
            measures difficult to apply
DUC-02 similar to 01, but
     single summary reflecting author view
     multiple summary as report
  some systems producing 'semi-extracts'
DUC-03 similar to 02, but
     single summary very short
     multiple geared to event/viewpoint/question
  evaluation intrinsic on quality                                semi-purpose
                          coverage                               quasi-purpose
              extrinsic on usefulness on source value            pseudo-purpose
                          responsiveness to question             pseudo-purpose
  coverage low, usefulness, responsiveness fair
DUC-04 similar to 03, with
     single summary as headline
     multiple for events, questions
       also English summaries for translated Arabic sources
  evaluation intrinsic on quality                                semi-purpose
                          coverage (mainly ngram similarity)     quasi-purpose
              extrinsic on responsiveness to questions           pseudo-purpose
results still baseline <= systems < humans

(Road Map 2)
DUC-05 :
    short multiple document summaries
    user-oriented questions, style (generic/specific)
  evaluation (with multiple human summaries)
             intrinsic on quality                                semi-purpose
                          coverage (ngram)                       quasi-purpose
             extrinsic on responsiveness                         pseudo-purpose
  hybrid systems, statistical + symbolic (parsing)
  results still baseline <= systems < humans
DUC-06, same as DUC-05, but also
  intrinsic evaluation on coverage by nugget pyramids
```


Figure 7: Summary of DUC evaluations

		Evaluation			
		'intrinsic'			'extrinsic'
		semi-purpose	quasi-purpose		pseudo-purpose
		quality	coverage		
			nugget	ngram	
DUC 2003					
Task					
1	single-doc very short	x	x		x
2	multi-doc short event	x	x		
3	" " " viewpoint	x	x		
4	" " " question	x	x		x
DUC 2004					
Task					
1	single-doc very short	x		x	
2	" " short	x	x	x	
3	multi-doc ex Arab very short			x	
4	" " " " short			x	
5	" " short	x	x		x

Figure 8: Details of DUC tasks, evaluations DUC 2003-2004

```
title: American tobacco companies overseas
narrative: In the early 1990s, American tobacco companies
tried to expand their business overseas. What did these
companies do or try to do and where? How did their parent
companies fare?
granularity: specific
```

Figure 9: DUC 2005 topic for summary