Technical Report

UCAM-CL-TR-679
ISSN 1476-2986

Number 679

UNIVERSITY OF CAMBRIDGE

Computer Laboratory

Automatic summarising: a review and discussion of the state of the art

Karen Spärck Jones

January 2007

15 JJ Thomson Avenue
Cambridge CB3 0FD
United Kingdom
phone +44 1223 763500

http://www.cl.cam.ac.uk/

© 2007 Karen Spärck Jones

Technical reports published by the University of Cambridge
Computer Laboratory are freely available via the Internet:

http://www.cl.cam.ac.uk/techreports/

ISSN 1476-2986

Abstract

This paper reviews research on automatic summarising over the last decade. This period has seen a rapid growth of work in the area stimulated by technology and by several system evaluation programmes. The review makes use of several frameworks to organise the review, for summarising, for systems, for the task factors affecting summarising, and for evaluation design and practice.

The review considers the evaluation strategies that have been applied to summarising and the issues they raise, and the major summary evaluation programmes. It examines the input, purpose and output factors that have been investigated in summarising research in the last decade, and discusses the classes of strategy, both extractive and non-extractive, that have been explored, illustrating the range of systems that have been built. This analysis of strategies is amplified by accounts of specific exemplar systems.

The conclusions drawn from the review are that automatic summarisation research has made valuable progress in the last decade, with some practically useful approaches, better evaluation, and more understanding of the task. However as the review also makes clear, summarising systems are often poorly motivated in relation to the factors affecting summaries, and evaluation needs to be taken significantly further so as to engage with the purposes for which summaries are intended and the contexts in which they are used.

A reduced version of this report, entitled 'Automatic summarising: the state of the art' will appear in *Information Processing and Management*, 2007.

Automatic summarising: a review and discussion of the state of the art

Karen Spärck Jones

1 Introduction

In the last decade there has been a surge of interest in automatic summarising. This paper reviews salient notions and developments, and seeks to assess the state of the art for this challenging natural language information processing (NLIP) task. The review shows that some useful summarising for various purposes can already be done but also, not surprisingly, that there is a huge amount more to do.

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.

The paper is organised as follows. The remainder of the Introduction notes the stimuli to summarising research in the last decade. Section 2 presents basic frameworks for characterising summarising systems, for evaluation in general, and for summary evaluation, that are used in the sections that follow. Section 3 considers summary evaluation in more detail, and analyses the evaluations that have been done so far. Section 4 examines the coverage of factors affecting summarising in systems and tests so far. Section 5 reviews implemented system design classes, with exemplar illustrations. Finally Section 6 offers an assessment of overall progress in understanding both summary task requirements and in building systems to meet these.

The Dagstuhl Seminar in 1993 (Endres-Niggemeyer et al. 1993) represented a first community attempt to promote research on a task that had, apart from scattered efforts, seemed too hard to attempt. The 1997 ACL Workshop (*ACL-97*) can be seen as a definite starting point for major research effort. Since then there has been a rapid growth of work on automatic summarising, worldwide, illustrated by a large literature including two books (Mani and Maybury 1999; Mani 2001). This research has been fostered by many workshops and

further encouraged by the Document Understanding Conferences (DUCs), now in their sixth cycle (*DUC*). The DUC programme, actually, despite its name, about summarising, owes much to the style and lessons of the Text REtrieval Conferences (TRECs - see Voorhees and Harman 2005). It has addressed the very difficult issues of summary evaluation through road maps designed to specify versions of the task, and performance criteria for these, in a way that is realistic given the state of the art at any time, but promotes a coherent advance.

Research on summarising since the mid-90s has been driven not only by ideas going back to the beginning of automatic summarising in Luhn's work (Luhn 1958)), but also by the general development of statistical approaches to NLIP, as illustrated by Language Modelling, and by successes with hybrid symbolic and statistical approaches to other complex NLIP tasks like information extraction (IE) and question answering (QA). The more recent QA evaluations within TREC have included questions seeking extended answers that are a form of summary, and teams participating in the DUC programme have also been participants in the earlier IE evaluations in the Message Understanding Conferences (MUC) programme (see Chinchor 1998) and in the QA evaluations. The performance levels reached with statistical and hybrid techniques in other areas, though not always high, have been sufficiently respectable to suggest that they offer a practical approach to useful summarising, where more ambitious strategies that exploit semantic and discourse information, of the kind discussed at Dagstuhl, can only be long-term goals. The general improvements in NLP technology, for example in fast and robust parsing (Appelt et al. 1993), and the arrival of solid public tools, like part-of-speech taggers, have made it much easier to put together experimental rigs for exploring new tasks and strategies for tackling them.

At the same time, the huge growth in digital material, and especially full text, has naturally stimulated a demand for systems that can produce derivatives that are highly concentrated on particular themes and topics, whether by selecting particularly informative initial text material, or by producing wholly new text to a more compact form, or by some combination of the two. This explosion of digital material has occurred in both the public and the non-public domain, but with different consequences for summarising work.

Much of the material in the non-public domain, for example proprietary journal databases with subscription access, is of the kind with which the original automatic summarising work was concerned; but the difficulty of obtaining open test material, and the proprietors' focus on other concerns, together with the technical opacity of much journal material (e.g. in chemistry) have meant that more recent summarising research has in general not tackled this type of text material. It is equally difficult to get test collections for other non-public material like enterprise data, which are often heterogeneous enough to be a different kind of challenge for summarising; and enterprise systems for managing these data have also concentrated more on improving other facilities like indexing, categorisation and search. The public material, on the other hand, and in particular the news material from which recent test collections have been predominantly drawn, presents its own distinctive challenges and opportunities for summarising, especially through the extensive repetition of text content: for example, this repetition may make it easier to identify salient content and ensure that even quite coarse summarising techniques will pick up anything important from one or another similar source.

This flood of digital text, and notably open Web text, has thus been a stimulus to work on summarising in multiple ways. It has encouraged a demand for summarising, including summarising for material in different languages and in multi-media contexts, i.e. for speech as well as 'born' text and for language associated with image data. It has also emphasised the multiple roles that summarising can have, i.e. the different forms that summarising as

an NLIP task can take, even if the classically dominant roles, namely prejudge or preview vis-a-vis a larger source text remain important: this is closely associated with the browsing, 'cut-and-paste' model of information management that IT has encouraged. The text flood has at the same time made it easier to develop or enhance NLIP strategies that rely on statistical data about language use.

The demand for automatic summarising has been matched by the NLIP research community's confidence (or at any rate belief) that, compared with twenty years ago, they are much better equipped with techniques and tools and by their experience with information extraction in particular, to make a non-derisory attack on summarising. Potential clients, notably the 'military-security complex', have reactively, as well as proactively, raised their performance expectations. In a more general way, the rampant march of the Web and Web engines have encouraged the perception that NLIP can do amazing things, and thus the expectation that new and more powerful facilities will come on stream all the time: the summary snippets that engines now offer with search results are normally extremely crudely extracted, but this does not mean they are not useful, and they illustrate the continuously improving facilities that the engines offer.

For all of these reasons, the status, and state, of automatic summarising has been transformed in the last decade. Thus even though most of the work done has been on shallow rather than deep techniques, the summaries produced have been defective in many ways, and progress in relation to 'quality' summarising has been very limited, something has been learnt about the task, a good deal has been learnt about some summarising needs and some summarising technologies, and a useful start has been made on coherent experimental work in the field.

2 Discussion framework

As indicated, this review will consider both the character of the summarising technques and systems as have been explored in the last decade, and such task performance results that have been obtained in evaluation studies. Since the work reported has been very varied, I will exploit some earlier description schemes as ways of analysing approaches to summarising and of examining system performance. (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

As a very general framework for characterising summarising systems I will use that presented in Sparck Jones (1999). This defines a summary, taking text as the classic though not essential form of input and output as

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

Sparck Jones (1999) then assumes a tripartite processing model distinguishing three stages, as shown in Figure 1: *interpretation* of the source text to obtain a source representation, *transformation* of the source representation to obtain a summary representation, and *generation*

of the summary text.

Definition and framework seem obvious, but are deliberately intended to allow for more variety in what constitutes a summary and in how it is derived than is now too frequently assumed. Thus the definition refers both a summary's conceptual content and its linguistic expression, and the framework allows for both minimal surface processing, transferring some given source text to summary text, and much more radical, deeper operations that create and transform meaning representations. The amount of work done at the different stages can also vary greatly, not merely between but within systems. Much current work is focused, under the label *extractive* summarising, with various approaches to surface processing, for instance by choosing different source-text selection functions; but where such extractive summarising seems to be inadequate for some summary purpose, a shift to *abstractive* summarising is proposed. This is intended to identify and re-present source content, following what is taken to be the generic style of conventional abstracts, as for academic papers, i.e. to be informative rather than indicative, to use the same language as the source, perhaps a similar content ordering, etc. However there are other summary forms that include digests and reviews, and range from query-oriented quotations to populated template population, which satisfy the definition and to which the framework can be applied, for both analysis and comparison purposes.

Summarising factors

The simple framework of Figure 1 applies just to summarising systems in themselves, as processors. But summarising systems are not context free. It is essential, as discussed in Sparck Jones (1999) and further developed in Sparck Jones (2001), to make the task for which summarising is intended explicit: there is no natural or best summary of a source regardless of what summarising is for. As Endres-Niggemeyer (1998) for example makes clear, professionals develop summaries on the basis of knowing what they are for. The design, and evaluation, of a summarising system has therefore to be related to three classes of *context factor*, as shown in a condensed and slightly modified version of Figures 2-4 of Sparck Jones (2001) in Figure 2. These are the *input* factors that characterise properties of the source material, e.g. language, style, units, etc; the *purpose* factors that bear on summaries, including their intended use and audience; and the choices for *output* factors, like degree of reduction and format, that depend on the the input and purpose features of any particular summarising case. (The sketchy factor characterisation given in the figure will be filled out using specific system examples in later sections.) There is no point in comparing the mechanisms used in different systems without regard for the summarising purposes for which they are intended and the nature of the source material to which these mechanisms are being applied. Equally, there is no legitimacy in system assessment for output without regard to purpose and input data. For proper evaluation, of course, the purpose factors have to be solidly enough specified to ground the actual evaluation methodology used.

Evaluation elements and levels

There are however many choices to be made in designing and conducting an evaluation. These can be developed using the decomposition framework for evaluation developed in Sparck Jones and Galliers (1996). This covers the evaluation *remit* and the evaluation *design* intended to meet this remit. Thus, as shown in Figure 3, the remit has to establish the evaluation motivation and goal, and a set of choices collectively determining what may be labelled

the nature of the evaluation. The evaluation design then has to locate, or position, the summarising system appropriately in relation to the remit. The input and purpose factors of Figure 2 define the *environment variables*, along with the output factors insofar as their generic character is clearly implied by the system's purpose or, indeed is explicitly stated, perhaps even in detail. The *system parameters* and their settings reflect the processor structure of Figure 1. To complete the evaluation design, this view of the system in relation to the evaluation remit has to be filled out with choices of performance criteria, evaluation data, and evaluation procedure. Again, the discussion of actual evaluations will fill out the brief evaluation characterisation given in Figure 3.

One feature of evaluations has been particularly important for language processing tasks, and has figured in summary evaluation. This is the distinction between *intrinsic* evaluation, where a system is assessed with respect to its own declared objectives, and *extrinsic* evaluation, where a system is assessed by how well it functions in relation to its encompassing *setup*. Thus for example summaries may be intrinsically evaluated against a system objective of delivering well-formed discourse, and extrinsically against a setup requirement for summaries that can replace full scientific articles for information reviews for busy researchers. Experience with summary evaluation since Sparck Jonea and Galliers (1996) has suggested a finer granularity is needed, from *semi-* through *quasi-* and *pseudo-* to *full-purpose* evaluation as shown in Figure 4 and further discussed below. This emphasises the point that evaluation without any reference to purpose is of extremely limited value, and it is more sensible to think of a continuum from the more intrinsic to the more extrinsic. Thus even an apparently intrinsic assessment of text well-formedness presupposes well-formedness is required in the task context. The gradations in the figure may seem over-refined, but as the discussion later illustrates, they are grounded in experience.

As summarising overall is so rich and complicated, I will use what has been done in summary evaluation as a route into my analysis of systems work. Thus I will take the way researchers have tackled evaluation as a way of addressing what summarising is all about, considering first evaluation over the last decade in the next section and then, in the following section, the major approaches to summarising that have figured in more recent research on automated summarising. Though this strategy is the reverse of the more conventional one which begins by considering summarising models and then how successfully they have been applied, it may supply a better picture of the state of the art.

The context for summary evaluation has also been influenced by the development of NLIP system evaluation in general in the last fifteen years. Evaluation methodology and practice has been seriously addressed for different NLIP tasks, notably speech transcription, translation, information extraction, text retrieval and question answering. These tasks variously share characteristics and technologies. This has encouraged transfers of evaluation notions from one to another, including ones from earlier-addressed tasks like translation to later ones like summarising. These transfers are not always well taken, specifically by failing to distinguish evaluation against system objectives from evaluation against larger setup purposes, so meeting objectives is taken to mean satisfying purposes. Experience with summary evaluation in the last decade shows the distance is greater than earlier believed.

3 Summary evaluation

Some of the earlier research on automatic summarising included evaluation, sometimes of a fairly informal kind for single systems (e.g. Pollock and Zamora 1975), sometimes more organised (e.g. Edmundson 1969; Earl 1970), also with comparisons between variant systems (e.g. Edmundson 1969) or against baselines (Brandow et al. 1995). The growth of interest in summarising during the nineties prompted the more ambitious SUMMAC cross-system evaluation (*SUMMAC* 1998, Mani et al. 2002). The DUC programme (*DUC*) in turn represents a more sustained effort to evaluate summarising systems. It has been of value directly in providing information about the capabilities of the systems tested, with some additions from other tests using the DUC materials though not formally part of the programme. But it has been of more value so far in forcing researchers in the field to pay attention to the realities of evaluation for such a complex NLIP task, both in terms of how concepts like intrinsic and extrinsic evaluation are to be interpreted and hence how useful they are, and of how sufficiently detailed evaluation designs can be formulated.

The DUC programme's original, and revised, road maps envisaged a progression from essentially internal system-oriented evaluation to external purpose-oriented evaluation. But the challenge of devising a true task-oriented evaluation for summarising, engaging with the *contextual* task for which summaries are to be used has, not surprisingly, proved far more difficult than for other NLIP tasks where evaluation programmes have been able, in one way or another, to limit evaluation scope. Thus for speech recognition, for example, evaluation has normally has been limited to transcription, and for retrieval to a system's ability to deliver relevant documents, especially at high ranks, in both cases ignoring larger task interests on the basis that doing better with these core components automatically assists the larger task. Information extraction has de facto followed a similar core-focused strategy.

It is much harder to pin down a summarising core component, certainly in a form which offers system developers much insight into its parameters or from which useful predictions about larger task performance can be made. This difficulty has been compounded by the fact that the researchers involved have come from very different summarising starting points and by the fact that, on the measures so far applied, system performance has been far from high. This makes it hard to develop task-oriented evaluations that that are both related to researchers' interests and are not too far beyond their systems' capabilities.

The DUC programme, along with some related programmes, has nevertheless played a significant role in encouraging work on summarising. The next sections review the major evaluation concepts, over the intrinsic/extrinsic spectrum, i.e. from least to most involved with purpose, that have been deployed in summary evaluation, and consider DUC and other evaluation programmes.

Summary evaluation concepts

The problems of summary evaluation, and some common evaluation strategies, already figure in Pollock and Zamora (1975). Much of what has been done in the last decade can be seen as an attempt to firm up and, as importantly, to scale up, earlier approaches and to move from the kind of approach used in Edmundson (1969), which explicitly eschewed any evaluation for the literature screening purpose for which such summaries were intended, to task effectiveness testing.

In the earlier work on summarising, it was evident, first, that producing coherent dis-

course was in itself an NLP challenge: thus the sentences in a Luhn (1958) abstract, however individually grammatical, did not when concatenated give a coherent summary text, syntactically, semantically, or referentially. Moreover even if sentence-extractive approaches like Luhn's do in general deliver syntactically well-formed sentences, there is no good reason to limit automatic summarising to these methods, and there is therefore a general summarising requirement to produce both well-formed sentences and well-formed discourse. It was also evident, second, that capturing key source concepts for a summary is hard, given we are dealing with complex conceptual structure, even on some 'simple' *reflective* view of a summary as a source writ small for the source text readers' preview. It was further evident, third, that measuring success in coherent delivery and concept capture is a tough problem, again even on some simple reflective view of the source-summary relationship.

Text quality

There is no absolute requirement that summarising output must consist of running text: it can consist of, e.g., a sequences of phrases, or a tabular format with phrasal fillers. But the need to produce running text is sufficiently common that it seems reasonable to start evaluation simply by considering whether the system can produce 'proper' sentences and 'properly connected' discourse. NLP has advanced sufficiently to produce both proper sentences and locally cohesive, even globally coherent, discourse. Thus for this kind of 'preliminary filtering' evalution it is appropriate to apply a series of text quality questions or checks, e.g. 'It should be easy to identify what the pronouns and noun phrases in the summary are referring to.' It may well be the case in practice that users in some particular contexts can tolerate a good deal of ill-formedness, but text quality evaluation is still valuable, especially for system developers, and it has played a significant role in DUC.

However quality questions are easiest to answer for local phenomena, within individual sentences or between adjacent ones. When they refer to a summary as a whole they are bound to be either restricted to specific phenomena, e.g. anaphoric references, or rather impressionistic. In particular, it may be hard to establish true semantic coherence for technical subject matter. Of course summaries may be misleadingly coherent, e.g. suggesting links between entities that do not hold in the source, but even human summaries can be defective in this.

Unfortunately, as Marcu and Gerber (2001) point out, quality assessment is too weak to be a system discriminator. The more substantive point about text quality evaluation, however, is that is does in fact, even if only in a low-key way, refer to summary purposes: the system's objective, to produce well-formed discourse, or even just phrases, is geared to what this output is for. The convention that refers to text quality assessment as intrinsic evaluation should relabel it as the *semi-purpose* evaluation of Figure 4, and recognise this in any evaluation detail.

Concept capture

The second question, does the summary capture the key (appropriate) concepts of the source, is much harder to answer or, more particularly, to answer to measurable effect. Even for a 'plain' reflective version of summarising, establishing that a summary has this relationship to its source is extremely challenging, not only because it involves judgements about conceptual importance in the source but because concepts, especially complex relational ones, are not clear cut and they may be variably expressed. We may wish to specify precisely what

should appear in the summary, but this is impossible in other than unusually constrained contexts. In general, asking for important source concept markup while leaving open how this should appear in the summary is too vague to support evaluation by direct source-summary pairing. Trying to control the process leads naturally to the model summary evaluation strategy considered below.

The basic problem is that humans do not agree about what is important in a source. As Rath et al. (1961) showed, even when given the relatively restricted requirement to pick out a specific number of the most representative sentences in a source, agreement between his human subjects was low. It would in principle be possible to handle this, and hence evaluation, via a degrees-of-agreement strategy, but multiple source markup would be very costly, and there is still a problem about whether the markup specification could be made sufficiently robust, without being unduly prescriptive, to support useful system evaluation by direct source (markup)-summary comparison.

The literature for and on professional abstracters (e.g. Rowley 1982; Endres-Niggemeyer 1998) suggests that important source content markup is a key practical process, but only as one element in the summarising whole. Using source markup as a basis for evaluation is thus problematic for this reason regardless of the others. It nevertheless seems that proper summary evaluation should consider the relation between summaries and their sources. Some other evaluation methods have therefore been used which refer to this, albeit indirectly rather than indirectly.

Edmundson's (1969) and Brandow et al.'s (1995) checking of summaries for acceptability against source was a weak procedure. But there are problems with the tighter or more focused form of summary-source comparison that reading comprehension tests appear to offer. Morris et al. (1992) used standard education assessment comprehension tests to investigate the effects of summary condensation, but does not discuss the implications of the type of question used. Minel et al. (1997) and Teufel (2001) use more sophisticated questions referring to source argument structure. However using questions about significant source points that a summary should also be able to answer, as in SUMMAC (*SUMMAC* 1998, Mani et al. 2002), is bound to be somewhat hit-and-miss where rich sources are concerned, and Kolluru and Gotoh's (2005) experiment is too small to support their claim that the method is robust despite human subjectivity. More generally, as Minel et al. point out, this strategy again involves an implicit reference to context and purpose: just as the notion of reflective summary implies that this is the sort of summary that is required for some reason, the same applies, in sharper form, to the reading comprehension model. This point is addressed in Farzinder and Lapalme (2005)'s lawyer-oriented questions. But more generally, reading comprehension is a typically underspecified variety of *quasi-purpose* evaluation.

In general, therefore, direct source-summary comparison has not figured largely in summary evaluation. It seems a plausible strategy in principle. But it is methodologically unsound when divorced from knowledge of summary purpose which could mandate source content that should appear in any summary. The main reasons it has not been used in practice, however, appear rather to be the effort involved except in the weaker versions just considered, rather than recognition of its methodological weakness.

Gold standards

In practice, direct source-summary pairing has been replaced by the use of human reference, or *gold standard*, summaries, so comparison for agreement on significant source content

can be considered without the complication introduced by the source-summary condensation. Most of the automatic summary evaluation for content capture done in the last decade has been on this more restricted, summary-summary pairing basis. It can be applied rather straightforwardly to extracted sentences, when the human summarisers are instructed that this is the form of summary required, and also to the more usual form of newly-written summary through a content '*nugget*' markup process. With non-extractive summaries human assessors are still required, both to do the reference summary markup and to judge whether, and how far, the reference nuggets are captured in the system summaries (c.f. the SEE program used in DUC (Over and Yen 2004)).

Unfortunately different human beings do not agree even on what sentences to extract to form a summary, let alone write identical or, often, very similar summaries, especially where there are no heavily constraining summarising requirements, e.g. specifying precisely which types of information are to be given, as noted in Rath et al. (1961)'s study when viewed as gold-standard extractive summary evaluation, and considered recently by, e.g., Daumé and Marcu (2004) and Harman and Over (2004). Thus model summary variations may swamp system variations (McKeown et al. 2001), and comparisons between reference and system summaries are likely to show many differences, but without any indication of how far these affect summary utility for the end-user. One way round this, as with source-summary pairing, is to have multiple human summaries, with the reference extracted sentences, or nuggets, ranked by their mutually agreed status, as in the Pyramid scheme (Passonneau et al. 2006). However this increases the evaluation effort, especially when the need for many reference summaries to counteract the effects of variation is fully recognised (van Halteren and Teufel 2003). Moreover where human aseessors are required, as with nugget comparison, variation can be large (Lin and Hovy 2002b; Daumé and Marcu 2004; Harman and Over 2004), which implies many judges are needed for evaluation to be really useful.

One advantage, in principle, of the gold-standard strategy is that appropriate deference to summary purpose can be built in. Thus as long as the human summarisers write summaries for the specified use and take account of other purpose factors like those shown in Figure 2, evaluating automatic summaries by comparison with human summaries ought to indicate the automatic summaries' fitness for the purpose in question. However there are two problems with this. The first (as painfully learnt in document indexing) is that human output is not necessarily well-fitted to purpose. The second, more important, point is that alternative outputs may in fact be as, or even more, satisfactory for the purpose in question. It is true that while unobvious index terms may work well for autonomous system searches, summaries have to be comprehensible to people. However this still allows for summaries very different from given human ones, and for effective summaries that do not fit closely even with the most agreed human content. (Fitness for purpose also applies in principle to the earlier source-markup strategy, but is even harder to manage than in the comparatively 'packaged' reference summary case.)

With extractive summaries in particular, automatic comparison between reference and system summaries is perfectly feasible, and the technology for ngram comparison, originally applied to machine translation, has been been developed in the ROUGE program and applied to summary evaluation (c.f. *ROUGE*, Lin 2004). It can allow for multiple reference summaries, and indeed for evaluation against other system summaries. It can also be used to compare non-extractive summaries, though clearly lexical agreement is likely to be lower. As a technique for evaluating summaries it is however much less informative than for translations, since with translations it is quite reasonable to bound comparisons, e.g. by sentences,

or at any rate to expect local rather than global variation. With whole-summary comparisons more variation can be expected, so similarity is likely to be much lower. As the method is applicable not just to individual words but to strings, it can implicitly take some account of well-formedness and not just lexical similarity, but only (de facto) in a limited way. It is thus evident that ROUGE-style evaluation for summarising is a very coarse mode of performance assessment except for specific, tightly defined summary requirements. Proposals have been made for more sophisticated forms of automatic comparison designed to capture cohesion, by Hori et al. (2004), or concept structures via graphs, by Santos et al. (2004), but these do not escape the fundamental problems about the gold standard evaluation model. There is the problem of model summary variation and, as Daumé and Marcu, and Harman and Over, point out, of variation in human assessors in e.g. identifying nuggets or comparing them. The implication is that multiple measures of 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.

In this spirit, Amigo et al. (2005) put forward a more ambitious gold standard methodology, using probabilistic techniques to assess, choose among, or combine different similarity metrics for comparing automatic and model summaries. But as they acknowledge, it all depends on having satisfactory gold standard summaries (though possible many alternative ones), and there has to be independent validation for the gold standards. The gold standard model therefore, however inadequate, is thus more correctly labelled quasi-purpose evaluation, as in Figure 4 than, as usually hitherto, as intrinsic evaluation; and as with the previous evaluation concepts the real status of the method deserves more examination in any particular case: specifically, what is the assumed purpose and what grounds are there for supposing the gold standard summaries, especially newly written rather than existing ones, satisfy it.

The foregoing implies there are relatively early limits to what can be learnt about summary merit independent of demonstrated utility in a task context. System developers may indeed find any of the forms of evaluation mentioned extremely useful in helping them to get some idea of whether their systems are doing the sort of thing they would like them to do, but encouraging system developers in their beliefs is not the same as showing their beliefs are well-grounded. However, as the specific evaluation examples described in Sparck Jones (2001) and also Sparck Jones and Galliers (1996) imply, genuine purpose-based evaluation for some task systems is bound to be extremely expensive, and it is natural for those trying to build systems with any significant capabilities to start by working only with much cheaper and simpler test protocols. Moreover given some widely available data and evaluation conventions, it is natural to adopt a 'suck it and see' approach, trying new ideas out using existing evaluation rigs, particularly since this allows obvious comparisons with others' systems as well as reducing costs. This has been long-established practice with text retrieval. But it has the major disadvantage that it emphasises mechanics, and diverts attention from the contextual conditions within which the original data collection and evaluation apparatus were based and which should properly be reassessed as appropriate for new systems. (This 'suck it and see' strategy is of course quite different from 'suck it and see' out there with real users, considered further below.)

There are other problems with the forms of assessment just considered, which have not been sufficiently recognised. One is evaluation scale. Though DUC and other programmes have increased test scale, summary evaluation has generally been modest in scale, and in some cases very limited in all respects. As Jing et al. (1998) show, evaluation may be very sensitive to specific data and context conditions, for example required summary length. Though the

range of environment variable values and system parameter settings covered in comparisons has slowly increased, sensitivity analysis is still too rare.

Baselines and benchmarks

Gold-standard summaries, specifically manual ones, have been taken as defining a target level for automatic summarising systems. Direct comparisons with them do not themselves define upper bound task performance, but they may be used to obtain a target task performance level, just as in retrieval the performance obtained with careful, manually formulated search queries is de facto a target for automatic systems. It has also become standard practice to define a *baseline* level of summary satisfactoriness or task performance, that any automatic system worth its salt ought to outdo.

One such baseline, for extractive summarising, as been *random* sentence selection, and indeed any system that cannot do better than this has problems. Perhaps more instructive baselines for news material have been taken, as in Brandow et al. (1995) and later in DUC, as a suitable length of opening, *lead*, source text. This baseline strategy depends on the particular properties of news, where sources are typically opened with a summary sentence or two. It will not necessarily work for other kinds of source, and it would be useful to develop a more generally-applicable form of baseline, or rather *benchmark*, analogous to that given by the basic '$tf * idf$'-type weighting with stemmed terms in retrieval: with sensible interpretations of $tf * idf$ this gives respectable performance. Thua one possibility for summarising would be a 'basic Luhn' using sentences ranked by a similar form of weighting. This could be justified as a simple, but motivated, approach to automatic summarising that it ought to be possible, but is not trivial, to outdo; and indeed this form of benchmark has been used in practice. It is motivated, in particular, as delivering summaries that could have some practical utility in task contexts. The strategy could also be applied, with suitable adjustment, to multi-document as well as single-document summarising. However since $tf * idf$ weighing varies in detail, it could be useful for researchers to adopt a common standard as a benchmark.

With some particular summarising strategies it may be possible to define upper bound performance (Lin and Hovy 2003), but such upper bounds, while useful, cannot be of general application.

Recognising purpose

The need to address evaluation conditions explicitly and afresh, and to cross what has been taken as the intrinsic/extrinsic boundary, is illustrated in Figure 5, which shows a reduced version of Sparck Jones (2001). This instantiates the evaluation specification shown in Figure 3 for the particular case where police reports about potential road obstructions from sleeping animals are the base for alerting summaries to a local population. The evaluation is intended, by questioning the townspeople, to show whether the summaries published in the local newspaper have been more effective as warnings than mobile police radio vans. It is evident that there are many other possible ways of establishing the alerts' effectiveness, and equally that different evaluations would be required to show, for instance, whether alerts with graphics were more effective than simple text alone, and also, whatever these evaluations might seem to show about the summarising system's effectiveness, whether police time spent on sending round warning vans and attending to accidents due to this type of obstruction was actually reduced.

At the same time, the original police reports might be taken as source material for a quite different purpose, namely as information sources for a biological researchers' database, for which a different sort of summary would be required and, of course, quite different evaluation.

This example about wombats may appear frivolous. But all of the elements have serious analogues: thus police reports about traffic were used to produce alerts in the POETIC project (Evans et al. 1995). Summaries, rather than just extracts, as potential material for a database are illustrated by *BMJ* (the *British Medical Journal*). Here 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 telegraphese (see Figure 6). Questionnaires are a standard evaluation device. The examples also emphasise, crucially for the present context, the point that though both the alerting evaluations can be labelled extrinsic ones, they illustrate different levels of evaluation from Figure 4. Thus the questionnaire-based evaluation is a form of *pseudo-purpose* evaluation, addressing the real summary audience but asking about putative influences on their driving behaviour rather than finding out about their actual driving. The police time analysis, on the other hand, if done with a suitable before-the-alerts and after-the alerts comparison, embeds the police activities that are responses to the alerts' causes within a larger setup evaluation. This is thus a *full-purpose* evaluation.

Research on automatic summarising so far has, as noted, touched on only a few and scattered choices among the many output possibilities, though they have been given larger scope by suppressing context detail and assuming that same-size-fits-many is a respectable summarising strategy. But the examples in Figure 5 imply that it is important, before taking over existing evaluation datasets and performance measures, to check that their context implications for a system are recognised and acceptable. Thus even for two cases which are quite similar, namely producing alerting summaries from input incident reports as done in POETIC and imagined in Figure 5, it would be easier to assess summaries for the former for factual accuracy and appropriate timing, because derived from particular traffic accidents, than the latter, which are generalisations.

The ramifications of context are well illustrated by the real *BMJ* case. *BMJ* offers two types of summary, for different materials. Editorials and news items are summarised using lead-text extracts, while research papers have formatted abstracts with subsidiary mini-abstracts - which are sometimes phrasal or telegraphic in style - per field. These differences are partly attributable to differences of the sources themselves, but much more to the purposes that the summaries are intended to serve. They presumably reflect different combinations of readership and reader interest, i.e. use, including multiple uses which may apply both to the extracts and the abstracts. In this they also illustrate the fact that even a single notional use e.g., say, scanning for general background knowledge, has many variations: thus different researchers may note, on the one hand, that yet another study has been done on condition C, another researcher that this is a study on C in the elderly.

Purpose evaluations

The DUC programme has had proper purpose evaluation as an eventual goal (*DUC*). More generally, it has been recognised that it is necessary to address the task for which summarising is intended (e.g. Hand 1997), not least because, as Okurowski et al. (2000) make clear, what happens in real world situations introduces desiderata and complexities that make focusing on the summarising system per se a recipe for inadequate or inappropriate systems, as well

as wasted effort.

Thus while summary evaluation so far has mostly been of the kinds already considered, some purposes have been envisaged, though without any associated evaluation, and there have been serious purpose-oriented evaluations. These have normally, however, not been *full-purpose* evaluations, but only *pseudo-purpose* ones, with varying degrees of simplification of or abstraction from full working environments.

The main summary use considered so far has has been for relevance filtering for full documents in retrieval. This was assumed, for example, 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 tests all used a protocol where relevance assessments on summaries are compared with those on their full sources. This seems simple, but is methodologically challenging. Thus comparing subjects' assessments on summaries with reference assessments of sources, (i.e. against what Dorr et al. (2005) call gold-standard *annotations*), improperly changes the people involved. However it may be difficult to avoid untoward priming effects if the same users judge both sources and summaries. Dorr et al.'s experiments avoided this but for a rather special document set situation. As is the case with retrieval testing generally, large samples are needed for stable comparative results when user assessments are individual. Earlier tests used existing rather than task-tailored summaries, but query-independent ones. Tombros et al. (1998) compared these with dynamic query-biased summaries, to the latter's advcantage.

Overall, there have been enough evaluations, both individual and within programmes, of summaries envisaged as or designed for full-document screening to support, despite the limitations of individual evaluations, some general conclusions about summarising for this use. Thus just for this one component function in information seeking, it appears that summaries are respectable predictors of source relevance. But it is also the case that very different summaries are equally effective, because the task is not a demanding one, so performance in this task cannot be taken as a useful indicator for others, and the retrieval task as a convenient evaluation proxy for other tasks.

Retrieval is an independently existing task with an established evaluation protocol. Other purposes for which summarising systems have been intended are clearly legitimate in that we observe that they exist in real life, though there is no established test protocol (or handy data) for them, and we can see that they either already involve summarising or summarising might benefit from it, say by digesting masses of material. However just as with retrieval, as a task that can take many different forms depending on local context, other tasks can come in many guises and may indeed be just one functional element in a larger whole as, for example, question answering, so the particular evaluation task has an artificial independence about it. This line may be pushed further when the evaluation task is based on a hypothesised function for summaries. Thus Minel et al. (1997) evaluated summaries as potential support for writing syntheses of sources.

Summary roles related to the retrieval one that have had some evaluation include support for browsing, to preempt or facilitate source review. Browsing is hard to evaluate: Miike et al. (1994) reports a simple time-based evaluation. This group of purposes includes making question answering more efficient, as in Hirao et al. (2001). Summarising has also been used as an internal system module to improve document indexing and retrieval, as in Strzalkowski et al. (1998), Sakai and Sparck Jones (2001), Lam-Adelsina and Jones (2001) and Wasson (2002); however here evaluation is by standard retrieval methods.

The other major class of purposes evaluated so far has been report generation, whether

these are viewed as digests or briefings. However there have not been many purpose-oriented evaluations here. McKeown et al. (1998) report only an informal user-oriented study on the value of patient-oriented medical literature summaries, but one in a real application context. Jordan et al. (2004)'s evaluation for data-derived summary briefings in a similar clinical setting is yet more embedded in a real context, as well as being a more substantive evaluation. In McKeown et al.'s (2005) evaluation of Newsblaster, report writing was used as means of evaluating summaries as sources of facts: system-generated reports were not the subject of evaluation.

Summary evaluation programmes

The DUC evaluations

DUC has been the first sustained evaluation programme for automatic summarising. But as it has been considered in detail elsewhere (e.g. in Over's overviews), I shall consider only its salient features here, focusing on what we can learn about the general state of automatic summarising from it and regretfully ignoring the the substantial and instructive detail to be found in its workshop proceedings and website (*DUC*).

The programme was based on a broad road map (Road Map 1) that envisaged a gradual advance from less to more challenging summarising along major dimensions: for input, from monolingual to translingual, single document to multi-document, unspecialised material, like news, to technical material; for purpose, from 'generic' or general-purpose reflective summaries to specific-purpose ones typically requiring more transformation, of one kind or another, of the source material, or longer rather than very brief summaries (and hence informative rather than indicative ones); and for output, as moving from summaries where fully cohesive text might not be mandatory to properly integrated and coherent ones. As importantly, it was envisaged that evaluation would progress from the less demanding, but sufficient for the early phases, intrinsic evaluation to, eventually, serious task-based extrinsic evaluation.

In fact, changes were made even in the initial stages: thus multi-document summarising figured from the beginning, largely because participants were already working on it, and because it may not, for the reason indicated earlier, be harder than single-document summarising. On the other hand it proved difficult to move from news source material, through a lack of participant interest and background resources. Moreover formulating and implementing satisfactory evaluation protocols proved to be extremely difficult. Thus the programme has so far had two stages: an initial one from 2000 onwards, covering DUC 2001 - DUC 2004, and the second from 2004 onwards. The main features of the DUC cycles to date are shown in Figure 7. The mode of evaluation as labelled within the programme itself, intrinsic vs extrinsic, is shown on the left, with annotations reflecting the finer granularity proposed in Figure 4 on the right. It should be noted that within the DUC literature the term "task" refers to the specific different summarising requirements, e.g. produce short headling summaries of less than 12 words, rather than to task in the rather broader sense used in this paper.

The first phase, following the first road map, was devoted to news material, with evaluation primarily for output text quality and by comparison with human reference summaries, but with a slightly more overt reference to envisaged system output use in 2003 and 2004. The specific evaluation tasks covered a wide range of summary lengths, and even cross-language summarising; and they introduced some modest variation on purpose factors through specified different summary uses. These were sometimes only implicit, as in asking for summaries

focused on events or from viewpoints, but were sometimes explicit, as in seeking summaries for utility in indicating source value, or as responsive to question topics. The participants explored a range of strategies, all of an essentially extractive character, but ranging from wholly statistical approaches to ones combining statistical and symbolic techniques, and with such hybrid methods applied to both source material selection and output generation. Particular groups sometimes used the same, or very similar methods for different DUC tasks, but sometimes quite distinct ones, for example Lite-GISTexter (Lacatusu et al. 2003) as opposed to GISTexter (Harabagiu and Lacatusu 2002). However as Figure 7 shows, the results in all four cycles up to 2004, while better than the first-sentence(s) baseline, were consistently inferior to the human reference summaries. In particular, coverage of reference content was low.

But this would not necessarily mean that the automatic summaries were of no utility for specific purposes. The problem with the initial moves towards task-oriented evaluation attempted in the various styles of summary sought in these evaluations in 2003 and 2004 was that the constraints they imposed on summaries, for example creating summaries geared to a topic, were already rather weak, so evaluation via comparison with a supposedly appropriate human summary was very undemanding indeed. Thus while this version of quasi-evaluation was intended to be more taxing than comparison against general-purpose summaries, it was not noticeably discriminating. At the same time, the content-oriented nugget evaluation by human judges was both expensive and not unequivocal, while the ROUGE-style automatic evaluation was not very informative. Moreover the first attempts, in 2003 and 2004, to address summary task roles more explicitly by asking human judges about summary utility (as a guide to potential source value) or responsiveness to a prior question, in a kind of minimal pseudo-evaluation task, were not at all revealing. It is difficult to judge utility or responsiveness 'stone cold' in the absence of an actual working task context.

The difficulties and costs of the evaluations, as they became evident in DUC 2003, stimulated the emphasis on ROUGE as the means of coverage evaluation in DUC 2004, to see whether this fully automatic process could replace the semi-automatic nugget comparisons. The results showed fair correlation, but perhaps not surprisingly given the dominance of extractive approaches. It was certainly not clear that such a convenient mode of evaluation would serve as the sole useful one. There were also, in phase one as a whole, problems with the early limits of the text-quality questions for extractive summaries: they could function as a filter below which summary quality should not fall, rather than a first-rank evaluator.

Some of the complexities of the DUC programme are shown in Figure 8, which gives the tasks and their evaluation methods for DUC 2003 and DUC 2004 in more detail. The programme has sought to advance the state of art in tandem with appropriate measures of performance, but this has in in practice meant much more adhoc than systematic change, so that while it is possible to see some development in broad terms, it is almost impossible to make systematic comparisons. Thus while DUC 2004 attempted to tackle some of the problems that previous DUCs raised, it did not resolve them, and brought new ones with the work on ROUGE. It also brought new complexities by introducing wholly new tasks, namely deriving summaries from (mechanically or manually) translated Arabic source documents, as well as by other changes of detail. Overall, there was a general feeling that the evaluations were not providing enough leverage: thus cruder extractive methods performed poorly, but not significantly worse than rather more sophisticated ones, There were other causes of dissatisfaction too. Thus the focus on news, though valuable from some practical and funding points of view, meant that issues that other types of source present were never tackled.

All of these considerations led to a revised road map, Road Map 2, intended to move more

decisively towards serious purpose-based evaluation and to the first evaluation of this second DUC phase in 2005. This was much more tightly focused than previous DUCs, with a single task, creating short, multi-document summaries focused on a rather carefully-specified user topic with associated questions, and also at either at general or particular level, as illustrated in Figure 9. Evaluation was on the same lines as for DUC 2004, but to compensate for human vagaries, both the ROUGE-based coverage and responsiveness assessment was against multiple human reference summaries. These sets of human summaries were also used in a parallel study (Passsonneau et al. 2005) to see whether nugget-based evaluation could be improved by weighting nuggets by their human capture. At the same time, it was hoped that the more carefully developed user questions, as well as checking for responsive content rather than expression, would support more discriminating and also useful responsiveness evaluation.

In general, the more concentrated character of the DUC 2005 evaluation, with its methodology emphasis, was an advantage: for example the particular quality questions distinguished baseline from human from system summaries in different ways. Overall, however, for all of the evaluation methods. the general comparative performance continued as before, with many different systems performing the same, edging above the baseline but clearly inferior to humans. However it is worth noting that the ROUGE scores were strongly correlated with responsiveness performance: this could imply that with well- (i.e. purpose-) oriented reference summaries, a good deal might be learnt from quasi-purpose comparative evaluation, though this has to be qualified if absolute scores are low.

DUC 2006 continues the DUC 2005 model, with only minor modifications, apart from the official inclusion of the pyramid nugget-scoring mode of evaluation.

Overall, as inspection of the detailed results for DUC 2005 shows, while many systems perform equally well as members of indistinguishable blocks in relation to the performance measures used, systems that do relatively well on one measure tend to do so on another; however individual systems vary both for individual quality questions and for quality versus responsiveness. The lessons to be learnt for the systems themselves and for the types of approach to summarising they represent, are considered further in Section 6.

Other programmes

The second major evaluation programme for summarising has been the NTCIR one (*NTCIR*), over three cycles of Text Summarisation Challenge (TSC-1 - TSC-3) from NCTIR-2 in 2001 to NCTIR-4 in 2004 (NTCIR has other tasks as well). The programme in general resembled DUC, but with an institutionalised distinction between extracts and abstracts, implying different evaluation methods. However the programme involved explicit extrinsic task evaluation from the beginning, following SUMMAC models, i.e. pseudo-purpose evaluation, as well as intrinsic evaluation including both semi-purpose evaluation by text quality and quasi-purpose model comparisons/ The tests used Japanese news material, so lengths for abstracts were specified in characters. Like DUC, the details became more careful over time.

TSC-1 had three tasks, all single-document summarising. The first, extracting different numbers of important sentences, was evaluated by comparison with professional human extracting on Recall, Precision and F measures. The second, aimed at producing plain text summaries of different character lengths (abstracts in principle though they could be extracts), was evaluated in two ways, by word-stem comparisons with human summary vocabularies, and on content coverage against the source and readability. The third, aimed at producing

summaries as guides for retrieval, was evaluated against full sources in SUMMAC style. TSC-2 had two tasks: producing single-document summaries at different character lengths, and producing short or long multi-document summaries. The evaluation here used the same coverage and readability assessment as in TSC-1 for both single and multi-document summaries, and also a 'degree of revision' measure for the single-document summaries. There was no form of extrinsic evaluation. TSC-3 was the most careful evaluation, for example with source documents marked up both for important sentences for extracts, and useful sentences providing matter for abstracts. The tests were on multi-document summarising, again treating extracts and abstracts separately. The former were intrinsically assessed for Precision and for coverage, the latter for coverage and readability . The abstracts were also evaluated extrinsically, following another SUMMAC model, for question answering, but in modified "pseudo-QA" form checking only for presence of an "answer" using string matching and edit distance.

In considering the results, it is most useful to take those for TSC-3. Performance for the extracts showed low coverage (implying uneliminated redundancy) with middling Precision, with many systems performing similarly, though for this data (news material but presumably with different conventions in Japanese media), noticeably better than the lead baseline. The content evaluation scores for abstracts were low, with human abstracts much better, and generally low scores for readability, though with variations for the many different specific questions. The "pseudo-QA" scores look rather better numerically than the content ones, but it is difficult to determine their real significance and informativeness.

The NTCIR programme was similar in many ways to the earlier DUCs, in design, results, and also in the dominance of essentially extractive systems. Thus the tests used the same kind of material and reproduced the single-document/multi-document modes of summarising at various lengths. The evaluations were primarily intrinsic, including both semi-purpose quality assessment and quasi-purpose comparisons with manual summaries (presumably of a reflective kind). The results similarly exhibit relatively low performance levels, inferior to manual summaries. The systems were typically variations on sentence extraction using statistical and location criteria, perhaps with light parsing so subordinate material could be eliminated or, in multi-document summarising, to make it easier to compare sentences for similar and hence redundant material.

The topic-with-questions model for summarising adopted for DUC 2005 and 2006 clearly has a close relationship with one of the forms of question answering studied in the TREC Question Answering (QA) track (Voorhees 2005a, 2005b). Thus the QA tests have included ones for so-called definition questions, like "Who is X?" or "What is a Y?", which appear to seek, or at any rate justify, rather more extensive responses than factoid questions like "How long is the Mississippi River?": in the tests the responses were treated as a set of information nuggets. In a later development taking questions in series, similar responses were supplied to supplement the answers to earlier specific questions in the series. Evaluation depended on assessors formulating, partly a priori and partly a posteriori, a set of appropriate nuggets and, more specifically, on identifying some of these as vital. System performance could then be assessed for returning (a) vital and (b) acceptable nuggets. A similar strategy was adopted for the series response case.

These extended question responses do fall under the general heading of summaries. But the specific task differs from the DUC one in that there is no prior specification of the source, or set of sources, to be summarised; the documents from which a set of nuggets is drawn need have no relationship with one another in the way that the members of DUC set of multiple

documents have a broad content connection. There has also been no requirement in TREC for any kind of cohesive discourse structure in the output 'summaries'. Finally, the mode of evaluation is of a very narrow gold-standard type, continuing the generic model instituted for QA within TREC, that assumes that it is possible to establish appropriate responses to the specific input questions regardless of larger task context. In general, therefore, the TREC QA efforts do not throw much light on summarising as a whole or on its evaluation, though the techniques applied by participants in TREC QA have also been applied to DUC topic-oriented summarising (see DUC 2005).

Assessment of evaluations

Overall, in the evaluation programmes and DUC specifically, we can see, along with some definite progress in summarising technology, that summary evaluation is more complex than it originally appeared to be. A simple dichotomy between intrinsic and extrinsic evaluation is too crude and, by comparison with other NLIP tasks, evaluation at the intrinsic end of the range of possibilities is of limited value. The forms of gold-standard quasi-evaluation that are manifestly useful for other tasks like speech transcription, or machine translation and to some, though a lesser, extent for information extraction or question answering, are less indicative of potential functional value for summaries than in these cases. At the same time it is difficult even with such apparently fine-grained forms of evaluation as nugget comparisons, when given the often complex systems involved, to attribute particular performance effects to particular system features or to discriminate among systems. All this makes potential task performance in context extremely problematic. The Catch-22 situation is well displayed in Lin and Hovy (2003): they attribute poor system performance (for extractive summarising) to human gold standard disagreement, so humans ought to agree more. But attempting to specify summarising requirements so as to achieve this may be as much misconceived as impossible. The same issue arises with Marcu (1999b)'s development of test corpora from existing source summary data.

Outside the programmes, summary evaluation is increasing, but is primarily intrinsic. There has been very little purpose-driven evaluation, and even taking both programmes and other efforts together, there is no really substantial data on what makes summarising work for types of situation. Radev et al. (2003) report a very substantial comparative evaluation across multiple systems and using a range of measures, both against gold standards and with a limited form of retrieval task evaluation. The gold-standard comparisons suggest that more sophisticated measures that factor in inter-judge agreement are more satisfactory, but different measures rate systems differently. The results also show better performance for systems than baselines, but it is diificult to draw inferences about fitness for purpose from the results since even the retrieval measure seems to be viewed more as an abstract device for comparing summaries than as a serious pseudo-purpose evaluation.

However operational summarising systems have appeared in the last decade, notably as Web search engine components but also free-standing as with Microsoft Summariser or system in Newsblaster (*NWBL*), and we may reasonably assume these deliver results that people find useful. These systems are deliberately general-purpose, even the search engine ones and certainly Microsoft Summariser and Newsblaster. In such cases evaluation from just one purpose point of view may not be particularly valuable, while multi-purpose evaluation is a hard nut to crack. The fact that such systems exist does not, moreover, remove the need to evaluate summarising systems designed for particular task contexts both from a practical

performance point of view and to throw light on the relations between context factors, design choices, and system effectiveness.

An interesting intermediate case is reported in Moens and Dumortier (2000). Summaries of magazine articles were specifically designed to prompt source purchases. A proper task evaluation would establish whether sales increased. But in the meantime the commercial publisher was sufficient;y impressed by an informal quality comparison with the installed basic system to replace it by the authors'.

4 Factors explored

In Section 2 I introduced the three classes of factor, input, purpose, and output, affecting summarising and indicated how many individual possibilities these subsume, for example of source subject domain and text genre. This section considers the extent to which automatic summarising in the last decade has explored these factor possibilities, for instance the types of source text taken as input, or the uses for which summaries are intended. "Possibilities" here refers to the factors to which systems have been exposed. This does not necessarily imply that systems have been designed to respond to these factors, whether to compensate for awkward aspects, e.g. ill-formed speech input, or to exploit helpful ones, e.g. document section headings. Systems may be designed simply to digest whatever comes, though in this may implicitly take account of source properties, for instance differences in lexical repetition with genre. Some factors may seem to require explicit response, e.g. source language, though even here heavy statistical techniques may operate at abstract character or even word-image level as in Chen and Bloomberg (1998). Similarly, systems may not be tailored to specific purposes but offer a standard output assumed adequate for this purpose as for others, and deliver vanilla rather than specifically-geared output. Dealing with the three factor classes is not necessarily correlated in a straightforward way with the three stages of system processing. Thus while in general, interpretation responds to source features, these may also be taken into account by later stages; and similarily while output factors may most obviously affect generation, they may may influence transformation.

The subsections which follow review the factor values that have figured in summarising research in the last decade, with example references, and comment on the extent to which systems have explicitly responded to them. In the following section, 5, I consider system types and examples. Taking these together illustrates the choices of summarising strategy and system design that developers have adopted and their relationship to context factors, especially input and purpose ones. Have the system choices been explicitly intended to meet context requirements, or have they been mainly motivated by the state of the NLIP art and the hope that this will in fact satisfy the context requirements. Either way, what have the actual results been? In the final Section 6 I will attempt to draw together system designs and their performance consequences, insofar as evaluation evidence so far allows us to do this.

Input factors

First, input factors, as listed in Figure 2.

Form factors

The most important input factors are the source form factors. The dominant input type in recent summarising research has been news material, specifically news stories from agency text streams. There are obvious reasons for this. There are many potential customers for summarising systems working on news: the most obvious are 'intelligence analysis' agencies, whether governmental or commercial. There is a great deal of news material, which is more easily obtained than, e.g., journal article datasets. Working with news material does not require specialist vocabulary resources or extensive specialist subject knowledge. Moreover, since news material has featured conspicuously in the DUC programme, there are publicly available evaluation datasets which can be used to test ideas and measure performance against previous results. However other types of input have also been addressed, including technical articles, legal material, message data, and spoken dialogue, as detailed below.

Language

In relation to specific form factors, first language: most of the source that has been used has been in English, mainly as news but also, e.g., technical journal articles; experiments have also been done with Japanese, notably in the NTCIR programme, and Chinese (Chan et al. 2000), again as news material, with Dutch (Moens and Dumortier 2000) and German (Reithinger et al. 2000) for example, and with both raw Arabic (Douzidia and Lapalme 2004) and automatically translated Arabic news in DUC-2004. The DUC Arabic work may be labelled cross-language summarising: it resembles cross-language retrieval in requiring processing of material in language A to deliver output in language B, and presents similar strategy choices, namely either process then translate or vice versa. Summarising systems that deploy NLP resources clearly have have to respond to the source language, but with statistical methods this language-specific response may be no more than required for appropriate stemming, and otherwise language-neutral.

Register

Register refers to what may be called the linguistic style of the source as popular, scholarly, etc., that in principle needs response in summarising. Some text registers have been represented in work in the field, for example news as popular and technical articles as scholarly (e.g. Saggion and Lapalme 2000, 2002; Teufel 2001, Teufel and Moens 2002), also as legalese (Grover et al. 2003), and perhaps as email (Corston-Oliver et al. 2004) or technical chat (Zhou and Hovy 2005). More interesting forms of register have appeared with speech input, including lectures (Nobata et al. 2003) and presentations of various kinds (Furui 2005), dialogues (Zechner 2001, 2002) and meetings (Murray et al. 2005). But register does not appear to have figured as a recognised processing condition for source interpretation in recent summarising research, even for spoken material if statistical techniques are applied, though Zechner explicitly responds to such informal speech register phenomena as restarts, and Furui compacts extracted transcribed sentences to clean them up.

Medium

In relation to source medium, most input sources have started life as text though, as just noted, speech has also figured, when transcribed. Image and graphic material has also appeared. I am excluding image-to-image (or graphic-to-graphic, e.g. Futrelle 2004) summarising here, interesting though it is (see Li et al. 2001 and, e.g., Rother et al. 2006). However sources that combine image and language may call for summaries that combine image and language, as with video news where selected images are complemented by transcribed speech

phrases (Christel et al. 2002, Papernick and Hauptmann 2005). Graphic material may also be embedded in a text source and call for processing within overall source interpretation (e.g. Carberry et al. 2004). Non-text data, e.g. tables or other records, have also been taken as source material (e.g. Maybury 1995; McKeown et al. 1995, Jordan et al. 2004, Yu et al. in press). Explicit responses have been made in the speech case, for instance to provide sentence or speaker segmentation (Zechner 2004), and are obviously needed in the image and data cases just mentioned. Video summaries may be by offering key/representative images from series, but also on text captions (Toklu et al. 2000).

Structure

Considering text source structure, this is intended to refer to overt, explicitly marked structure, for instance headings, boxes, etc, rather than 'in-text' discourse structure as signalled by lexical data, whether as cue phrases or word frequencies or as embodied in syntactically and semantically expressed forms of text organisation like rhetorical patterns. This is not an absolute distinction: thus location (e.g. lead sentence, paragraph opening sentence) is on the border: but as associated with format may be taken as structurally explicit. News material, i.e. indiviual stories, does not have much explicit structure, apart from titles, and it normally follows the 'lead sentence' convention, to which summarising has directly responded. Quotations, which are common in news, are a form of marked item with a structurally-relevant emphasis or elaboration function, but one that has not generally been picked up in summarising strategies (though see Moens and Dumortier 2000). However citations in technical text, which have similar functions, have been used (Teufel 2001, Teufel and Moens 2002). Forms of overall explicit structure that have appeared include that for legal materials (Moens et al. 1997; Farzinder and Lapalme 2004; Grover et al. 2003), for magazine pieces (Moens and Dumortier 2000) and for technical journal articles exploited in McKeown et al. (1998) and Elhadad and McKeown (2001), and in Saggion and Lapalme (2000, 2002). Individual Web pages, taken as source in Radev et al. (2001a) may have complex structure, quite possibly with elements that have to be exluded in summarising (Berger and Mittal 2000b). Structures explicitly linking multiple items, in the form of Usenet threads, are used in Sato and Sato (1997). Speaker separation in dialogues, used in Zechner 2001, is perhaps a rather exotic form of structure marking.

Genre

Under genre, much news material represents a recognisably individual genre distinct from, say, narrative, which may be somewhat pleonastically labelled 'reportage', mixing event sequences with event and player property descriptions, though other genres like editorial comment figure. Some summarising systems, e.g. Newsblaster (McKeown et al. 2002), have been designed to respond to these specific reportage characteristics. Moens and Dumortier (2000) respond to opinions versus reportage for magazine articles, and Farzinder and Lapalme (2004) to 'direct' versus narrative in legal reports. Reportage overlaps both narrative and description as more general genres and these, together with argument, are characteristic of journal articles. Argument and description have been specifically addressed, for instance in the medical domain by McKeown et al. (1998), and by Teufel (2001) and Teufel and Moens (2002). Encyclopedia articles are a distinctively compact form of description, exploited as such by Salton et al (1997). One particular genre which may be deemed a subclass of instruction, as that represented by question-answer inquiry, treated by Sato and Sato (1998) and Zechner (2001), or more generally by negotiation (Reithinger et al. 2000, 2002).

Length

Finally, for the length form factor, news stories are generally relatively short sources, compared say with, e.g., technical articles. This has had a pervasive influence on summarisation research because it has permitted compression factors (e.g. 30%) which would be quite unrealistic for article summaries. However summarising work on long sources has usually finessed the problem of condensation by working with only selected source elements (e.g. 'Results' sections), or by delivering such brief summaries as headlines, or query-oriented snippets, that necessarily ignore most of the source, though Nakao (2000) explicitly addresses book summarisation. Multi-document summarising may have large input sets.

Other input factors

Subject

News material is, of course, very varied in subject content, and not usually opaquely technical though, e.g., financial and sports items are not a-technical. Thus while gazetteers are very useful, there is no call for extensive specialist subject lexicons. More generally, where technical material has been taken as source with, e.g., journal articles or Usenet, the summarising techniques adopted have not required any explicit recognition, or deep understanding, of technical content. But technical knowledge is used in extracting patient-specific material from medical papers in McKeown et al. (1998) and, more elaborately, for handling computer-related sources in Hahn and Reimer (1999). Apart from any implications about requirements for subject knowledge in summarising, or subject area associations with source text structures, the range of subject areas explored in summarising research so far as not been large, with law by far the most prominent (e.g. Wasson 2004).

Units

In relation to source units, one important development that working with news has stimulated is a quite new emphasis on summarising over multiple input units, i.e. on multi-document summarising, which was not previously a concern. Both earlier work on automatic summarising, and classical human summarising for, e.g., academic papers, has addressed single-document summarising. Moreover even where, for example, summary reviews have ranged over multiple source documents, these have not usually had the degree of overlap that is characteristic of newswire streams where new stories that update ongoing events are continually appearing with repeated background matter. The content redundancy in a set of news stories is typically far greater than in the linked passages studied by Salton et al. (1997), for example, and systems like those tested in DUC have been designed to exploit repetition across documents as an indicator of importance while removing it from output. Multi-unit summarising based on unit clustering has also been applied to Web pages (Radev et al 2001a), and summaries for individual units of the same type in different medical papers combined for output in McKeown et al. (1998) and Elhadad and McKeown (2001).

Authorship

Authorship as an input factor (e.g. single, multiple, known, etc.), though potentially important in other contexts is not usually important for news material. It does not appear to have been explicitly investigated for automated summarising, though it may be indirectly reflected in conflicting content from different input sources that it may be necessary or desirable to

indicate in summarising (McKeown et al. 1998).

Header

The final input factor is header, or more generally metadata, information, interpreted here as data assigned to sources rather than forming original parts of them and possibly consisting of data types outside the ordinary text language, e.g. classification codes. The boundary between this factor and structure is a loose one. It includes, for example, dates which with news may be publication dates not original story filing dates, reporter or news agency labels etc. Dates may be exploited in multi-document summarising to order extracted material. Zhang et al. (2003) explore the use of annotations, as they might be made by later source readers with a highlighter, for summarising. On somewhat similar lines, Sun et al. (2005) use dynamic clickthrough data for Web pages to adjust word values for sentence extraction.

It is evident from this detail that source properties have implications for summarising strategies: one important question is what the work that has been done so far says about the extent to which reasonable summarising performance can be obtained using general techniques, perhaps with some light-to-moderate tuning or tailoring to source characteristics, as opposed to systems designed primarily for a particular type of input material. (This refers primarily to deliberate modification with perhaps some statistical training: systems based entirely on machine learning would presumably have to be trained from scratch for distinct situations.) But it may be too soon to judge how effective general-purpose, or largely general-purpose, approaches may be, partly because most summarising strategies have been crude ones incapable of responding to many input features, and partly because those purposes for which automatic summaries have been evaluated have not demanded fine-grained or comprehensive input interpretation.

Purpose factors

Use

The most important purpose factor shown in Figure 2 is the use for which summaries are intended. This is the major factor in determining the content of the output summary and its presentation, as illustrated in Sparck Jones (2001). There are a number of generic use classes including supporting source preview, assisting source scanning, filtering sources, alerting on source content, acting as briefing substitutes for sources, and so on. These generic classes are quite broad: for instance a summary that is taken as a substitute for a full source may be a digest giving a reduced version of the whole source or an extract dealing with a particular element of the source, and so on. Summaries may or may not be responses to input queries. Uses may merge, for example preview and filter; scanning abstracts to find out what's happening may be mixed up with deciding to read particular sources.

This breadth and fluidity makes it difficult to exploit uses to guide summarising, or as a base for evaluation. Uses can, however, and often need to, be made more specific, for instance filtering on the basis of likely overall source interest for a topic versus filtering as likely to give the answer to a specific question, which might call for event-based or person-based summaries of news respectively. Again, purposes become more specific when related to different customers, e.g. briefing for specialists vs briefing for school children. There is also a global distinction between uses where users are people and where users are other systems or system components. Thus whether or not an extracted passage (one simple form of summary)

is helpful as the source of an answer to a question may be quite different for a human user and for a subsequent system answer-extraction module, and the same applies if summary information is directly offered to a reader or is used to populate an automatically-searched database. In general, automatic summarising work has assumed a human user.

Much of the summarising work done so far has made no explicit reference to summary use. There are several reasons for this. One is that the 'source reflection' model of summarising, along with evaluation against human reference summaries, is taken to imply that automatic summaries that compare well with human ones will therefore serve whatever purpose the latter are intended to serve, so it is not necessary to pay attention to what this purpose actually is. A second reason is that source-reflective summaries are seen as offering a neutral, condensed but content-covering, version of the source that is naturally suited to many, especially unknown or weakly specified, purposes. This position is further justified by the fact that existing summaries, e.g. of academic papers, are seen to serve many functions. The same point applies to summaries provided in response to queries, as in Web engines: the immediate function of the summary is addressed, but not the contextual use to which the summary is put. A third reason is that, except for some application-specific cases (e.g. Moens and Dumortier 2000), automatic systems have often not done well enough in capturing and presenting significant source content to demand more than direct and rather basic quality assessment, whether completely informal or by some kind of 'semantic nugget' checking.

However recent summarising experience has prompted a move towards evaluation geared to use, as in recent DUCs. One reason for this is the fact that general-purpose reflective summarising does not impose clear constraints on summary properties so there are no strong guides to system design. Another is the fact that different summaries or summary types perform equally well on measures that are not linked tightly to specified uses. When automatic summaries look rather inadequate to the human eye, especially when compared to those humans produce, there is a natural desire for more leverage in system building. There may also be a further reason for more focus on use: systems are appearing that produce summaries of types not often encountered in human summarising, like Google's snippets, so there are no existing models to refer to.

Summarising research in the last decade has therefore included work addressing system uses. It has been stimulated both by the evaluation programmes and by particular applications. Several of the use types mentioned have figured in this work, though sometimes only in very particular forms. Many papers refer, in very vague terms, to potential summary uses. In some cases, system design has been motivated by envisaged uses but has not reached task-based evaluation. In other cases there has been some task-oriented evaluation.

The main form of use has been support for document retrieval, and in particular support for relevance judgements without reference to full sources. This is a role for summaries going back to the earliest automated summarising work, and it has featured in evaluations in SUMMAC, DUC, and NTCIR as well as in, for example, Brandow et al. (1995), Tombros et al. (1998), and Dorr et al. (2005). The carefully-limited retrieval use considered in these evaluations is a manageable and satisfiable one for automatic summaries. But it has proved a poor discriminator between methods, or guide to what more might be needed for other tasks. Related uses that have been taken as motivation for summarising have been quick 'overview' scanning of system output for document sets (e.g. Strzalkowski et al. 1999), or skimming for individual documents (Boguraev and Kennedy 1999) or, more generally, browsing where, for example, summaries may be generated on the fly for 'current' sections of long documents, as in Miike et al. (1994). Two other, interestingly specialised forms of this use

are represented by support for the disabled, in audio telegraphese for scanning (Grefenstette (1998), and subtitling for the deaf as a sort of gisting for skimming (Vandegehinste and Pan 2004). The need to consider the implications of particular uses for what summaries provide is well illustrated by Moens and Dumortier (2000), where highlighting summaries are specifically designed to encourage browing users to buy their source articles.

The other main type of task addressed has been briefing. This has been mainly in application-specific contexts and forms, for example summaries of medical information bearing on patients (McKeown et al. 1998, Jordan et al. 2004), and of 'to-do' lists (Corston-Oliver et al. 2004), but also in more open and varied forms in Mani et al. (2000), though none with reported evaluation. Definition question answering, as investigated in TREC (Voorhees 2005b) can be seen as a form of briefing. Support for report writing as envisaged in Minel et al. (1997), and taken as a way of evaluating Newsblaster summaries in McKeown et al. (2005), is a closely related use. More generally, the question/topic-oriented summarising tested in DUC can be seen, if not directly as briefing or report generation, as support for these, as also in Hirao et al. (2001).

Many papers handwave about possible uses, but these are apparently without impact on system design. Others seem to assume that as some form of manual summary, e.g. headlines, is common, it must have uses, so automatically generated headlines may be legitimately evaluated, i.e. retro-fitted, for a task like retrieval assessment, as in Dorr et al. (2005). More generally, post hoc evaluation has been adopted as a way of seeing whether some current or feasible summarising method can deliver results that in practice serve some purpose, without designing from scratch, or drastically modifying an existing system. The general problem with this approach is that it favours obvious strategies and obvious 'bug' fixing, rather than task-motivated analysis. It has been particularly common with extractive techniques and with retrieval as the favoured task, perhaps partly on the ground that the form of retrieval evaluation, through relevance judgement, is well-defined, and partly because extractive methods might be found adequate for such tasks. It is possible to suppose, however, that the snippet summaries offered by Web search engines are based not only on what is feasible but on the view that they should be helpful, though so minimal, in relevance assessment (broadly defined). But test conditions have often been set for output, notably length, on some presumption that there is a need for summaries of different lengths, without any particular contextual rationale through intended use.

Audience

In general, professional abstract writers for science sources, or their author surrogates, have assumed as an audience for their abstracts much like the readership for the full sources, i.e. informed scientists, and the same assumption is typically made for academic papers generally. This assumption refers both to the nature of the background knowledge that potential readers have, and to the nature of their interest in the material. However summarising technical scientific material for a popular audience is a familiar variation: indeed news articles about important papers are essentially summaries of this kind, perhaps mixing an account of the source with an assessment of its significance. Again, the familiar executive summary prefacing a report can be a subtle variation on the full source: thus while the full report may be targeted at managerial or political readers, an executive summary may be more more specifically targetted at senior executives who may not have the time or technical experience to fully absorb the whole source.

News material has at least two quite distinct audiences. One is indeed that for the full sources, which for many news sources is a broad and heterogeneous one with varied knowledge and concerns. However the other important audience is the professional intelligence analyst, where "intelligence" can refer to particular areas of interest like financial operations, or new technological developments, or criminal activities, or national security. Here the summary audience is assumed to be well-informed and also quite focused in the character and scope of their monitoring. Systems like Newsblaster address the first, general, news audience; the DUC programme, as it has moved towards extrinsic evaluation, has assumed the second, though in both cases the character of the audience has not obviously been factored into detailed system design. A wide and varied audience is assumed for Web page summarising in Radev et al. (2001a), and de facto in many retrieval-oriented evaluations.

Analysts are professionals. Professional audiences are assumed, if not specified in detail, for summaries of legal sources, as in Moens et al. (1997) and Farzinder and Lapalme (2004). Similarly, summarising for academic papers, as in Teufel (2001) and Teufel and Moens (2002), assumes the same technically informed audience as the sources do. Shared organisation or company membership also limits audiences, as for briefing summaries as in Corston-Oliver et al. (2004) Quite particular audiences, namely doctors in particular communities, are specified in McKeown et al. (1998), and in Jordan et al. (2004). A rather different take on audiences is embodied in Verbmobil's summaries (Reithinger et al. 2000), since the summary here is for the source originator(s), i.e. gives the essential data gathered from the extended negotiation in which they participated.

Such specific target audiences are taken as conditions in system design, or may be taken as corollaries of very particular uses, as in 'to-do' briefings, but in many cases target audiences have apparently not been taken as influences on design, other than indirectly as a byproduct of the source material: for example, the nature of the audience for summaries of some type of legal material may appear simply to follow from the nature of the material itself and so not need explicit attention in system design. The broader the audience, the more often proxies or notional representatives figure in evaluations: Very particular target audiences are more often explicitly invoked for evaluation.

Personalisation is a more particular form of audience recognition. This is of course automatic, but is analogous to query-oriented summarising. Zhang et al. (2003) consider the more complex form of personalisation represented by summarisation based on individual users' source annotations.

Envelope factors

As noted earlier, this group of factors covers a range of further requirements including time, e.g. for summary production or summary currency, target locations, formalities to be satisfied, e.g. legal constraints, summary triggering conditions, and destination, which may not always be a human reader, so is not synonymous with audience.

Time

In general, the time factors addressed in summarising to date have not been very tight, and have arisen fairly naturally from the nature of the material being summarised. For instance, since source news stories flow in continually, a system like Newsblaster is naturally led to timely updating in summary production, presumably by time-slice windowing, and perhaps also implying that previous summaries move gently into (possible accessible) history. The

traffic alert summaries in Evans et al. (1995) stem from a dynamic event database and clearly have to be delivered in a timely manner, and there would also be a time constraint in producing 'to-do' lists from briefings as in Corston-Oliver et al. (2004). Otherwise, there is a normal requirement for immediate summarising when the task is query-oriented summarising, as occurs in practice with Web engines and is assumed in query-oriented summary evaluation. Radev et al. (2001a) is unusual in considering time constraints from an engineering point of view when scaling systems up, for instance, to more users.

Location

Location has not been addressed in an very specific form, but is involved in the generic form of digital output to the user's workstation, especially in the context of Web browsers: thus Newsblaster (*NWBL*) explicitly offers clickthrough to, e.g., person-oriented summaries from event ones, and summaries may be linked with relevant images, as in Mani et al. (2000). Such linking may be part of the summarising system itself, or associated with the embedding system in the general style of modern information management (as in WebInEssence (Radev et al. 2001a) and MiTAP (Damianos et al. 2001). Even the simple idea of summary phrase highlighting (Boguraev et al. 1999) assumes the modern style of workstation in practice, in conjunction with scrolling, even if such highlighting could be displayed offline, and so do other ways of 'visualising' summaries, as in Corston-Olver et al. (2004), or of offering users ways of personalising summary presentations through different data 'views' (Aone et al. 1997). Location in a more and, in particular a constraining rather than enhancing, form is represented by output to handheld and hence limited-space devices, as in Boguraev et al. (2001) and Corston-Oliver (2001).

Formality

Formality refers to specific requirements that are not deducible from, e.g., use or audience, and often have a conventional character. Legal constraints about, e.g., summary vocabulary, or declared 'authorship', or liability disclaimers, may be one example; and conventions about the form of bibliographic data for sources in abstracting journals, or about the form and completeness of case citations for summaries of legal material are others. Using standard fixed headings for summaries, as in *BMJ* (see Figure 6) has something of formality about it, since though something like them is implicit in summaries given their intended use and audience, the particular set of headings is a formal condition. Summarising research has hitherto been so preoccupied with 'core' summary content that this factor has been ignored: for example research summaries for news often lack accompanying source attributions, though we may assume these would be supplied in operational systems as in Newsblaster (*NWBL*). Formal conditions may be easy to satisfy, as in this case, but in others are more difficult, e.g. assigning summary content to particular headings or avoiding statements that might be deemed libelous in a summary review.

Triggering

Specific triggering, is obviously involved in the normal requirement for immediate summarising when the task is query-oriented summarising, as occurs in practice with Web engines and is assumed in query oriented summary evaluation. Triggering is also to be expected for alerting summaries, as illustrated for the traffic case in Evans et al. (1995), where each arriving input triggers an explicit decision as to whether to issue an alert, responding either to a new incident or a material change for an existing one.

Destination

Finally, the default destination for summaries has been the human end user, and this has been reflected in, e.g., DUC evaluations, where the user's ability to supply an appropriate interpretation for perhaps ill-formed text is assumed. Clearly, while some of the styles of quality assessment used in, e.g., DUC may be useful here, what is offered the end-user has ultimately to be evaluated in a manner appropriate to the contextual task. This may include not only assessing content fitness but many other relevant properties from presentation ones like readability to general convenience or 'habitability' in the user's setup, which covers more than visual presentation. These destination conditions may interact with the summarising process itself, and not simply guide output factor choices. There do not appear to have been any significant overall evaluations from this angle, as opposed to specific ones on output. Where summaries are intended to serve other system modules, for example in passage extraction for question answering, the destination is the answer extraction module and this may imply not only specific design properties but a different form of evaluation. The immediate destination could also be a translation module (e.g. Douzidia and Lapalme 2004), which could in principle influence the nature of the summary, e.g. by requiring simple sentences since these are easier to translate, though this does not seem to have been explicitly investigated.

It will be evident from the foregoing that there are many purpose factors potentially affecting summarising that have hardly figured in summarising research so far, and more specifically have played little part in guiding system design. There are exceptions: thus the particular strategy developed for Lite-GISTexter in 2005 was motivated by the requirement for question-directed summaries (Lacatusu et al. 2005); again, the nature of some source data input to reporting, briefing or alerting has naturally led to summarising strategies that can deal with multi-unit material, as with news streams.

Overall, the purpose factor values investigated so far have been scattered and with only limited comparative systems. This makes it extremely difficult to reach any conclusions about whether there are generic strategies suited to generic purposes and, in particular, to generic uses. It may be that the collective effect of all the environment variables that define a task context is to make each case so individual that one cannot even assume that there is some type of summarising strategy suited to the summary use involved. As against this, we may perhaps argue that we have sufficient evidence to suggest extractive strategies are sufficient for a kind of retrieval use, and this could also hold for other generic uses. At least, more research is needed to determine this.

Output factors

As emphasised in Sparck Jones (1999, 2001), though input and purpose factors together impose the major constraints on summarising, they do not usually determine fine-grained choices about output properties, for example whether a formatted layout or running text is more effective for some summarising purpose where there is no presumption that source format has to be matched by summary format. 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 Zechner (2002)'s DIASUMM can produce both for spoken dialogues. In some cases, purpose factors may be quite definite and specific, as in formality conditions requiring a formatted layout with particular headings. But in general there are open options

for output where suitable choices can only be firmly established by evaluation in the task context, though in individual cases the existing situation may justify inferences about choices likely to suit purposes. It has nevertheless been the case that summarising research as a whole has ignored many output factors, mainly by assuming a reflective running text-in/running text-out approach. Output factor choices apply to both transformation and generation.

Material factors

Coverage

Material factors include the nature of the coverage of the source, i.e. whether comprehensive or selective. In general automatic summarising has assumed that all of the major concepts in the source will appear in the summary (subject to the size constraints on the summary). In multi-document summarising this is interpreted as referring to the concepts represented in the source set of documents, which may imply some loss of a particular document's angle on a common concept though its presence is represented. Reflective summaries are exactly those which are intended to be comprehensive. This is independent of technique: thus extractive summaries may still be intended to be comprehensive. Query- or, more broadly, topic- oriented summarising that in general does not cover sources comprehensively has been the main form of selective summarising explored so far, as in the DUC programme and TREC question answering, and also as implemented by Web engines. Summaries using particular headings may, on the other hand, not cover the source comprehensively, as illustrated by *BMJ* and by McKeown et al. (1998), Moens et al. (1997). Selectivity is natural with briefing or alerting, as illustrated by Corston-Oliver at al. (2004)'s action lists, and Evans et al. (1995)'s traffic alerts also ignored classes of source content.

Reduction

A material factor that has loomed unexpectedly large has been reduction, often called compression and defined mechanistically so the summary is expected to be, e.g., 10% of source length. Such mechanistic definitions, as adopted for several DUC cycles, are convenient as rough versions of softer constraints like "short", and as guiding operators on statistically-based extractive summarising processes, for example continue adding sentences until the set length is reached (regardless of content 'rounding off'. In practice, however, they have often not reduced sources very much (e.g. by 50%), and where sources are of variable length do not deliver comparable length summaries: this contrasts with publication conventions that call for, e.g., summaries of (about) 200 words regardless of source length. More generally, reducing short news stories to 30% of source length may give quite short summaries but not be very challenging for the summarising condensation process. The other major length specifications in experiments to date have been for fixed length, e.g. 100 words, as in DUC, or for the extremely short 'one sentence' or 'headline'-style summary, again as in DUC and e.g. Banko et al. (2000), Dorr et al. (2003) and Zhou and Hovy (2004). In many experiments to date these reduction factors have been investigated without much regard for their specific use justification, and rather as trials of systems' condensation capabilities within an overall 'reflective' framework, as for example in Grewal et al. (2003). More motivated work on reduction, both locally as telegraphese and globally has, however, been undertaken not only to support quick skimming, as with phrasal summaries (Oka and Ueda 2000) and for audio (Grefenstette 1998) or along with video (Christel et al. 2002), but also to fit summaries to limited spaces, as with handheld devices (Boguraev et al. 2001, Corston-Oliver 2001).

Derivation

The third material factor, derivation, refers to whether the summary text reproduces source text, at the clause or sentence level, i.e. more than just lexically, or re-expresses source content. Derivation has an independent rationale, but has also been made explicit as an experimental parameter for summarising methods, as in NTCIR. For some summarising purposes it may be not merely appropriate but necessary to provide an extractive summary, reproducing source text, as illustrated in minimal form by Web engine summary snippets. This is not a matter of output choice but rather follows from intended use, by showing how the retrieved item relates to the query terms. But more generally, much summarising research has sought to show that extractive approaches are sufficient for the use in hand. This might be regarded as an output factor choice: summaries derived from source text in the sense of replicating it are suited to the summary use; but this something of a retrospective justification for much research practice. In other cases summary uses imply that derivation must not be constrained in such a way, for example summarising that requires a particular writing style which is not necessarily that of the source author, as in Corston-Oliver et al. (2004)'s 'to-do' briefing summaries. There are apparent intermediate cases, for example where source sentences are truncated and perhaps tweaked, as in Harabagiu and Lacatusu (2002); but this is essentially derivation from source. There are potentially complex cases too, e.g. summaries of sources that combine source quotations with comment on them, which have not been significantly investigated so far.

Specialty

The fourth factor, speciality, refers to the implications that the purpose audience rather than use have for how source material appears in the summary. Thus, for example, highly specialised technical detail may be unsuited to a non-technical audience, so some transformation with respect to level of technicality is required. This appears not to have been specifically investigated.

Style

Output style is a loose, but well recognised, notion. The classic contrast is between informative and indicative summaries, but there are other possibilities, e.g. reviews, plotlines, etc. In many human summarising cases, there is a clear implication from the summarising use that the output will be informative, for instance saying what experimental results are, not just indicating that some experiments have been done (cf the *BMJ*, Figure 6. In automatic summarising, indicative rather than informative may be an emergent property of the summary as a whole: thus where extractive techniques are used it may well be the case that even where individual parts of the summary are informative, the overall effect is only indicative. However a use like facilitating source assessment in retrieval can be taken as a clear justification for indicative summaries, and specifically for phrasal summaries as in Oka and Ueda (2000), while Boguraev and Kennedy (1999)'s phrasal summaries are designed for source skimming. Summarising based on 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). Question-answering summaries, and particularly summaries in response to definition questions, have to be informative: this is not merely rational but evident in the nugget-based evaluation methodology used in e.g. DUC 2004. Saggion and

Lapalme (2002) offer both indicative and informative summaries, wtieh the latter amplifying the former.

Format factors

The final format class of output factor covers a set of sub-factors that may in some cases follow rather tightly from the purpose specification but in general do not.

Language
At the top level, the language used, e.g. English versus Japanese, is a direct consequence of summarising purpose and not a matter of choice, and may of course imply an output language that is different from the source one, as in DUC tests for English summaries of automatically translated Arabic. Just as with source material, nearly all summarising research has been with English output, with Japanese in the NTCIR programme, and some work in other languages. e.g. German (Reithinger et al. 2000).

Register
However there is also the choice of language register or linguistic style, e.g. plain and simple versus complex. In general summarising research has not made a deliberate choice here. Thus extractive summarising repeats the language register of the source, and extracted sentence simplification is driven by the desire to remove unwanted content rather than simplify expression. Even if summarising is not extractive, the default reflective approach implies that there is no deliberate change of linguistic style or register from source to summary. However phrasal summaries, as in Boguraev and Kennedy (2000), and extracted window-defined snippets, embody a deliberate decision about the acceptability, even utility, of 'telegraphese', which may be viewed as a distinctive register for, e.g., scanning purposes. This also applies to some compressed headline summaries, which indeed may be barely well-formed word strings (Witbrock and Mittal 1999). Grefenstette (1998)'s audio telegraphese is clearly purpose motivated. Sata and Sato (1998) extract simplification by rewriting would constitute a register shift to make answer summaries to questions easier to understand in an instruction context. There are other examples where output is generated from a deep content representation, as with the Verbmobil summaries (Reithinger et al. 2000), which have no close register relationship to their source dialogues.

Medium
In relation to the summary medium, as noted earlier some summarising research has explored non-text source material and the production of non-text summaries to match (e.g. Rother et al 2006; Futrelle 2004), as well as combined image and text summaries for video news (cf. Li et al. 2001; Papernick and Hauptmann 2005). Images may also have a role as illustrations accompanying text summaries. Thus Microsoft Summariser and Newsblaster *NWBL* shows images, just as news source does, implying a system need to choose appropriate images from a potentially large source set. Selected video images may analogously accompany transcribed speech or caption-based summaries (e.g. Merlino and Maybury 1999). Grefenstette (1998) and Carberry et al. (2004) envisage audio as the primary output medium, as appropriate for disabled users. However supplying actual speech to accompany, e.g., text extracts from transcriptions could be not only illustrative and amplifying through expressive detail, but also an insurance against poor transcription. There appears, however, to have been little

work, apart from Merlino and Maybury (1999)'s investigation, to explore media options, and especially text versus graphical or image output as alternatives for summary effectiveness, though Papernick and Hauptmann (2005) consider alternative presentations of image and text video summaries.

More generally, and taking a wider view of media, modern workstations as output devices offer a range of presentation facilities that can be used for summary 'visualisation'. These include, e.g., highlighting, as in Boguraev and Kennedy (1999), but they also make it possible to present multiple views of the summary and its context, and for the user to engage in dynamic interaction with the source and summary material, as illustrated by Aone et al. (1997), Ando et al. (2000), Mani et al. (2000) and Radev et al. (2001a). These complex possibilities make evaluation in a task situation essential, but also extremely difficult.

Structure

As noted earlier, the summarising purpose may specify a particular (explicit) structure under format, as in information extraction generally. However it may also be an option choice for output. Even quite elementary choices can affect perceived utility, e.g. if a summary is a list of phrases, should these be presented as a list, and in source or alphabetical order? Some research has adopted a particular output format, which may be well-suited to IE-style summarising strategies, for example using forms with headings and fillers as in Maynard et al. (2002) and, in deliberately tabular form, in Farzinder and Lapalme (2004); White and Cardie (2002) use their IE structure as the base for rich hypertext output, while Mani et al. (2000) use a script-like structure for organising multimedia biographical summaries, delivered as forms. But there has been no serious work on comparative value for a task. In general different views of the information 'resource' embodied in the source material together with different forms of summary from, e.g. extracted sentences to keyword lists, as in Merlino and Maybury (1999), are seen as complementary rather than competitive. Thus the complex interfaces and forms of visualisation just considered under medium (e.g. Radev et al. 2001a) represent more elaborate forms of output structure, not limited to the summary per se but embedding this in a larger body of structured information.

Genre

Finally, genre. Research on text generation has explored genre, and output genre figured in earlier summarising research, most noticeably where summarising is from non-text input so an explicit choice of text genre for the output is required, as in Maybury (1995)'s choice of report mode.. However the default reflective model of summarising and, more particularly, the dominance of extractive summarising, has implied that genre is not a specific output choice, especially for single-document summarising. As noted, since Newsblaster offers a choice of both event and person-oriented summaries, these carry with them, to some extent, narrative and descriptive genres respectively, though these also partly follow from the different system strategies used. Again, summaries as responses to questions for definitions may imply a descriptive genre, but in all of these cases genre has essentially been emergent from the news material and summarising methods used, rather than a system parameter setting. Similarly, the production of apparently descriptive-style summaries from extractive materials in Elhadad and McKeown (2001) is primarily a consequence of the way material is selected from the source rather than a pure consequence of the summary purpose. However Reithinger et al. (2000) illustrates a deliberate choice of descriptive output genre.

Factor lessons

As the foregoing implies, the many factors involved in summarising, their individual complexity, and the enormous number of possible factor combinations, mean that so far only a few groups of cases have been explored in any remotely systematic way. This has not encouraged an analytical approach to system design as a conscious response to factor conditions, or at least to all factor conditions as opposed to the obviously pertinent ones like intended summary use. Implemented systems may rather have been based on assuming de facto adequacy for factor circumstances without any pressing need to recognise and react to these explicitly or, perhaps sometimes, in the belief that they can be actively ignored. (This is setting aside technical feasibility as a limit on factor recognition, though it may well have played a part in practice.)

There are families of systems with much in common, notably variant extractive ones; and it can be argued that the fact that these extractive strategies, that have dominated DUC and other programmes as well as summarising research as a whole in the last decade, are not completely useless, confirms Luhn's (1958) belief that there is *something* of value in such a simple approach, and that this is because these techniques do, somehow, respond sufficiently to many factor conditions or constraints. The DUC programme and its relatives, and the further experiments that their materials have encouraged, have been an important beginning in systematic exploration of the summarising terrain. But these programmes have also served, through the detailed evaluation specifications they require, as much to raise new questions about what summarising is all about as to answer such old ones as "Can we produce something that looks (feels) like a summary?" The relatively limited factor ground that the programmes have covered at all systematically has also served to emphasise how scattered and ad hoc summarising research coverage of the factor field so far has been. Again, the challenges of conducting proper task evaluations even for what seem to be relatively undemanding cases, that DUC has experienced, have meant that many researchers have been able to put off the day when they have to tackle systems that will adequately meet more demanding needs, for example ones that lay source argument structures bare, or that are grounded in more contextual state than a current information query, or that generate output text with maximum referential clarity..

It must also be recognised that all of the factor headings are broad labels hiding great complexity and it may be extremely difficult, even with careful task evaluation, to determine precisely which specific properties of sources, requirements of purposes, and choices for output matter. This difficulty is compounded when summarising is one function within a multifunctional system as in Radev et al. (2001a) and Damianos et al. (2002), as is increasingly the case in practice with, for example Web engines.

5 Systems: approaches and structures

The growth of summarising research in the last decade has naturally stimulated work on different summarising strategies and system designs. Earlier research had already explored both shallow, essentially statistical approaches (as originally in Luhn 1958), and deep, symbolic approaches (as in Hahn and Reimer 1999), along with various intermediate strategies (s in Earl 1970). More recent work, prompted partly by the evaluation programmes, partly by the growing supply of training data and processing tools like parsers, and partly by the practical desire to respond to external task needs, has meant both that generic types of approach,

especially statistical ones as illustrated by NeATS (Lin and Hovy 2002a) and MEAD (Radev et al. 2001b), have been explored in much greater detail than before, and that new types of approach, especially those combining statistical and symbolic techniques, have been investigated, as illustrated by SUMMARIST (Hovy and Lin 1999) and 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 this mass of work in detail. In this section I will use the basic system structure of Figure 1 first, to consider the types of model that underlie current systems and second, as a framework for analysing some individual exemplar systems. In many cases, published system descriptions give architectural and processing details without reference to underpinning models of summarising and their justification, and the nature of the intermediate representations used: this is particularly common for statistically-based approaches where the theoretical or at least practical motivation is taken for granted and the interest is in the fine variation, and where the notion of text *meaning* representation appears inappropriate. This recent development reflects practical and technological conditions, but contrasts with with earlier work focused on the nature of discourse structure and explicit text meaning representation. It is nevertheless useful to consider current approaches from the modelling point of view, in relation both to notions of what summaries and summarising in general are, and to how these are influenced by particular task conditions. Some distinctions, for instance between *extracts* and *abstracts*, or between *indicative* and *informative* summaries are often made, but these are extremely crude and fail to characterise individual systems and their applications properly.

Thus in considering the character of current systems, we should ask what kind of source representation they form and how they do this given their type(s) of input; what kind of summary representation they form and how this is derived from the source one, and what kind of output they deliver and how. Further, are the choices made driven by notions of what summaries ought to be like, relative the summarising needs, or by available technology options and post hoc practical validation?

For convenience I will group systems as extractive and non-extractive. Each, the former especially, covers many variations in current work, and there is also no absolute distinction between extractive and non-extractive. The recent interest in multi-document summarisation, which was not considered in earlier summarising research, complicates the picture. However the structure of Figure 1, originally seen as for single-unit input summarising, is in fact general enough to be applicable to the multi-unit condition. It is also the case that many systems have a complex structure with many modules that may not appear to fit the generic structure: the discussion which follows is primarily about logical system structure rather than the implementation module set or operational sequence. At the same time, while the output of the first interpretive stage constrains later ones, modules may be quite loosely coupled: this is evident in the way system variations are explored in DUC experiments, for example.

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 us 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.

Extractive strategies

Basic statistical approaches

It is natural to start with the simplest approaches, which happen to be the most commonly implemented, namely statistical ones.

The most basic version follows from Luhn, scoring 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: the source representations delivered from the first input interpretation stage for single documents is the minimalist one consisting of sentence identifiers in source order with the scores. The source sentences are not used to derive the summary representation in the subsequent transformation step. The summary representation is in turn just some selection of the sentence identifiers obtained after ranking by score down to a cutoff and then recovering the source order for the chosen items. If the cutoff is by a fixed number of sentences, matters are simple; however if it is by output summary word length, some sentence length information is also involved. The output generation stage then calls up the sentence texts corresponding to the representation identifiers. This is a logical view: in practice reordering may be done during generation. It It is also a purist view according to the abstract model and assuming a single requirements specification. In the alternative view the summary representation is the ranked set of selected identifiers with cutoff as well as reordering applied during generation. The advantage of this view is that it allows for different output length summaries from the same underlying representation using the ranking by scores.)

As content meaning representations both source and summary ones are very weak: the first implicitly indicates sentences' relative content importance as defined by lexical frequency data, and the second signals that certain sentences are especially important and thus summary-worthy. In both cases there are no marked relations between sentences. The fact that word frequency has something to do with corresponding concept importance motivates this strategy, and it can be argued that it is appropriate (as well as being easy to implement) for task contexts where some indication of source text content is sufficient. This can arise for a range of uses, audiences and envelopes.

But treating sentences independently, as the basic approach does, 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 includes lexical data, but it is logically an element of transformation.

Multi-document summarising, even at its simplest, leads to slightly richer processes and representations. In particular the source representation now normally includes some topic or theme identifiers, again derived using lexical statistics, so sentence scores are relative to the themes. The topics or themes are obtained by some clustering process, applied to whole document units or directly to sentences, with documents or sentences scored against some

cluster characterisation like a lexical centroid vector (e.g. Radev et al. 2000). A cluster of documents on a broader topic is usually taken as the basis 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 sentence words for future redundancy processing. Tranformation for the summary representation then involves the ranking, selection and ordering of individual sentences per cluster, and also coverage of different topic clusters. Transformation and generation are essentially similar to the single-document case, but with additional operations to factor in different subtopic, e.g. by applying a round-robin strategy or using metadata like source timestamps, and also, very commonly, to deal with sentence redundancy per topic, for example by applying MMR.

The motivation for cluster-type multi-document strategies is quite transparent: some applications have repetitive inputs, and while repetition can be taken to emphasise importance, it does not need to be carried over to a summary assumed already to be importance-based. At the same time it appears to be the case that multi-document summarising, even using simple statistical methods, requires more complex representations than single-document. Thus while it is in principle possible to treat all the pooled sentences as if they come from a single source, simple scoring on single-document lines may not be discriminating enough to select good summary sentences, or sentence comparisons over the whole set be sufficient to identify content repeats: clustering focuses both processes more effectively. With large document sets, moreover, subtopics may be more distinct and substantive than for single documents, and so deserve more recognition.

Enriched statistical approaches: lexical units and features

The generic strategy just outlined is clearly adaptable to, for example, query-oriented summarising through query term matching at some sentence-selection point. It is of course equally applicable to whatever is taken as a source unit or 'passage', for example to larger units like paragraphs, to subsentential units obtained by simple sentence segmentation, to text windows, etc. At the same time, it is naturally extensible to a more sophisticated treatment of the lexical elements for which statistics are computed, which includes both differentiating and differentially weighting element types and adopting particular type definitions for which complex identification processes are required. This extension may refer only to the interpretation stage, which delivers sentence scores and representations as before, or to approaches which include units in the representations so they are available for later operattions, 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 of development within the essentially statistical approach has been to use more varied and elaborate lexically-based features as the basis for computing sentence scores. This has included ones that are still statistically determined, using recurrent ngrams rather than words, or statistically-based multi-word elements like recurrent word pairs, or additional information obtained by applying statistical association techniques to identify topic signatures (Lin and Hovy 2000), i.e. sets of related words. More directly linguistic, symbolically-grounded tactics include invoking available lexical resources charac-

terising word senses and relations, like WordNet, or authority lists like gazetteers; using morphological or stemming operations to merge variant word forms; and, most importantly, 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 include topic or theme identification, are illustrated by Barzilay and Elhadad (1999) - see also Silbers and McCoy (2002), by Harabagiu and Lacatusu (2005) and by SUMMARIST (Hovy and Lin (1999) and Lite-GISTexter (Lacatusu et al. 2003). Current shallow, but robust, parsing techniques that exploit part-of-speech tagging can in particular be used to identify and select linguistically-significant multi-word lexical elements as sentence features, including those representing named entities and phrasal concepts like Filatova and Hatzivassiloglou (2004)'s 'atomic events'. This processing can also be applied to elements that are not part of ordinary language but may be important for summarising, like many proper names or other identifiers. Adding symbolic to statistical processing was first seriously investigated by Earl (1970), but modern tools offer far more possibilities and recent research has taken this line much further.

Specific lexical items with importance-signalling properties, e.g. "conclusion" in some domain literature, have also been investigated (e.g. Teufel and Moens 1997, 2002). So have other unit types with language-like properties, like Web links and URLs, which are more complex in detail but have many ordinary-language behavioural properties and have been exploited as text features, e.g. Chakrabarti et al. 2001. By natural extension, following early summarising research exemplified by Edmundson (1969) but now over a greater range of options, other forms of information, including metadata information, may be used as unit characterising and weighting features. Thus whether a sentence word also occurs in a document title, or with typographical emphasis, or a sentence in paragraph initial position, may give it extra weight.

The analysis processes used to identify units, whether statistical or symbolic, may be non-trivial; but the results may still only be used to derive sentence scores, or be carried forward simply as opaque characterising sentence features in the same manner as single words. They do not thus enrich the explicit form of intermediate representations to any marked extent, but remain 'bag of word' representations, albeit not completely trivial ones. The subsequent transformation and generation stages in summarising remain comparatively simple: overall, these statistically-based systems are quite elementary as processes which use and deliver discourse meaning.

Many of the systems built in the last decade have been based on the sentence extraction model. The main variation in this strategy is where subsentence units are reduced to phrases (e.g. noun groups) that are used directly as the representation elements for input sources rather than as features of other units, as in Witten et al (2000)'s simple keyphrase summaries. These are treated much as sentences in the transformation and generation stages, though they may undergo some additional operations, e.g. to choose the particular output expression for a set of morphologically variant phrases. The rationale for using phrase list output summaries is supplied by task applications like relevance filtering or browsing in information retrieval (as in Witten et al. 2000), where overall output summary text coherence may not be needed.

As noted earlier, many different systems have shown similar performance in the larger-scale evaluations like DUC and NTCIR. However it appears to be the case that where there are distinctions between better and less-well performing system sets, those that use more refined extraction procedures are often superior, and that the use of multi-word expressions, however identified, can be advantageous. This applies even to 'generic' summaries for news

and within the rather undemanding forms of evaluation mainly used so far. The situation may be different, and the gains sharper, in the context of other source types and summarising purposes. Thus one consequence of the query-oriented summarising of the more taxing kind tested in DUC 2005 has been a system requirement not merely for more sophisticated source sentence analysis but also for question analysis.

Enriched statistical approaches: structures

In relation to the larger range of strategy options, and summarising needs, two developments within the overall extractive approach are of particular importance. The first is a more comprehensive use of source structure.

Thus systems may use sentence structure characterisations not merely to identify units and features for scoring in the interpretation stage, but as 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 during the transformation stage, as in Newsblaster (*NWBL*, McKeown et al. 2002). The key condensation stage of summarising is thus more than operating on feature lists for source sentences: it involves various forms of structure comparison, merging, scoring and ranking.

But structure here is still sentence-level structure. There is no reference to (whole) source *discourse* structure (or dependent summary discourse structure) beyond the essentially statistical salience model for lexical units (and hence the concepts behind them) that motivates basic statistical summarising strategies, including cluster-based multi-document ones. But whether for single-document or multi-document summarising, such statistical models of structure, based on unit frequency and co-frequency, are still relatively simple ways of noting major elements and emphases in the source material, and do not necessarily capture any of the richer semantic or pragmatic structure that indubitably characterises discourse. However there are more complex but still statistical ways of capturing something about discourse structure beyond simple unit salience, as illustrated by Erkan and Radev (2004)'s use of graph structures based on sentence lexical relations to identify sentence centrality. 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, based on predicate-argument sentence analyses, to identify sentences to select for the summary, and Vanderwende et al. (2004) and Leskovec et al. (2005) also use graphs based on logical forms, for example ones expressed as triples.

Statistical approaches to structure determination, using lexical similarities, may however be used not only to identify different topics or themes, but to establish topic flow, particularly for single documents, which can be used to order extracts in output. Boguraev and Kennedy (1999), for example, divide source texts into successive topic segments using lexical overlaps. Summarising is per segment, and the segmentation is carried forward and used to organise the output summary. Nakao (2000) similarly uses lexically-based segmentation, but here at different levels of granularity forming a hierarchy, for book summarisation.

Other moves to identify and use semantic/pragmatic discourse structure, have exploited symbolically-defined structures, i.e. ones that address meaning explicitly rather than implicitly, as in the statistical case. In their simplest forms these approaches exploit rather weak forms of discourse coherence structure, as embodied in discourse protocols for focusing or

centring based on distinctions like given/new, and correlated anaphoric reference patterns. Boguraev and Kennedy (1999), for example, resolve anaphors within discourse segments as a means of improving counting information for substantive source content units. But attempts have also been made to use richer, and global, symbolic discourse structures. Most work has been done with Rhetorical Structure Theory (RST). Thus Miike et al. (1994) and Marcu (1999a, 2000) apply RST-motivated discourse parsing, exploiting discourse markers in particular, to build discourse structure trees, expressing types of text segment relationship, as source representations. These are then used, in Miike et al. by taking the relative importance of relations into account and in Marcu essentially by scoring tree node 'dominance' status, to identify key 'nucleus' source clauses for production as summaries. PALSUMM (Polanyi et al. 2004, Thione et al. 2004) build more abstract discourse structure using relations like subordination, and prune them to obtain summary text units. Teufel and Moens (1998, 2002) apply rhetorical discourse or argument categories rather than relations to identify important source sentences, also with the implication that, for the categories pertinent to scientific papers that they use, these would eventually be used to organise the output.

In both statistical and symbolic approaches of the kinds just mentioned, source and summary representations are usually of the same general type, with the latter some selection, perhaps reduction or radical simplification, of the former, obtained with varying transformational effort. Most of the system effort goes into capturing source structure rather than exploiting it. Thus in the RST case, the relations are primarily a means of building a nucleus-satellite tree through which key material can be identified, and they are not carried forward explicitly to supply the same type of structure for the summary.

However, as emphasised in earlier summarising work (e.g. Endres-Niggemeyer et al. 1995, Sparck Jones 1995), there are many possible types of generic discourse structure, and many variants of each: very broadly, linguistic, domain, and communicative types, each with top-down or bottom-up forms, where individual text structures either instantiate standard schemas or are constrcuted from standard relationships. There is further no reason to suppose that the structures motivating source and summary representations have to be of the same type, and there are good reasons to allow for summarising systems that make use of multiple structures, which may be deployed only internally in particular processing stages, or be manifest in representations. RST and the PALSUMM model are both linguistic models of a very 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 structures were of real use, especially as they cannot (with current methods) be identified very reliably. Moreover while such strategies as Marcu's might seem suited to particular contextual needs like selecting key clauses as headline summaries, it is less clear how to extract multiple sentences from entire trees in a way that delivers coherent longer summaries. Thus Marcu or Miike et al. can deliver multi-sentence summaries, but these also may be list-like rather than continuous text. But the main problem is that work with richer, specifically symbolic, discourse structures has been extremely limited and in many cases has not reached evaluation stage, e.g. for Carberry et al. (2004)'s use of structure defined by the communicative intentions, modelled as plans, behind in-text graphics.

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), especially ones when summarising verges on classical information extraction. With applications where 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 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 illustrate an input/response communicative structure. White and Cardie (2002) illustrate a full-blown IE-based approach well-suited to particular applications, where sources are analysed to fill template slots for world or domain as characterised by e.g. event types,

As well as illustrating the use of different discourse structure types, the work with symbolic structures also illustrates top-down model forms, as in McKeown et al.'s schemas, others bottom-up ones, as in RST; and individual systems may combine several types of structure. There is no reason, for example, to limit source structure analysis to a single type, and combinations might be more effective, as Mani et al. (1998) suggest; thus McKeown et al. and Elhadad and McKeown use linguistic features of the source, associated with the domain, to identify material for the domain-based source representation.

As noted, 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 the event templates obtained from source interpretation into sentence selection. However the (types of) source 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, syntactic centring and semantic lexical relationships. 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 reordered and reformulated (Elhadad and McKeown).

Harabagiu and Lacatusu (2005), illustrate both the possibilities and the complexities of working with discourse structure. But they show there is value to be gained (for multi-document summarising) from working with larger-scale discourse structure and specifically with general-purpose structural models rather than application-specific ones and, further, from exploiting several types of structure including explicit as well as implicit ones. Thus they go beyond statistically-based topic determination and topic segmentation to exploit syntactic, predicate-argument and semantic pattern information, using this to identify thematic structures that express source conceptual (i.e. domain) content. These may be represented in graph form and combined, through both shared content and linguistic discourse relations to form larger themes. Their approach is essentially bottom-up, using both linguistic and domain (world) information to identify key source content and select and order it for output.

Comments on extractive summarising

Many variations, less or more complicated, of the extractive approach have been explored in the last ten years. The problem in assessing the value of greater interpretive sophistication, whether in deeper sentence processing or use of above-sentence discourse structures, has been that the forms of evaluation in the multi-party evaluations have not been challenging and

discriminating enough, while those in more challenging task settings have not made sufficiently informative wide-ranging comparisons. Assessment is made more difficult by the fact that many systems are characterised more in terms of local parameter choices than by reference to substantively different summarising models.

The growth of summarising research has led to a wider range of comparators, not only baselines like 'lead' or random but also, as noted earlier, of a type of benchmark represented by some $tf * idf$-style weighted Luhnian approach. This is useful in offering a non-trivial comparator, and may focus attention on what more the user's task context requires, but does little to guide choices of direction for rather different approaches. Conscious comparisons between rather different approaches as in Harabagiu and Lacatusu (2005) are therefore especially instructive. Some may argue that just looking at what a summarising system actually does is all that is needed to understand and assess it, and that there is no need to characterise systems using abstract models and to fuss, for instance, about whether some particular process is done at interpretation or transformation stage. My argument is that a more careful model analysis is valuable in understanding the role of individual processes and in making it possible to relate process choices to the constituent conditions and requirements of any particular summarising task, with the further implication that systems can be flexibly and effectively parametrised for different situations.

Machine learning

This argument applies even though the second major recent development within the extractive approach has come with the introduction of machine learning, for example applying LSA or SVM techniques, and using both supervised and unsupervised methods.

Given the range of possible features for characterising source units, it is natural to ask whether modern machine learning techniques can be applied to determine which features are most useful for source unit characterisation to support summarising, and how they may be weighted and combined. (Kupiec et al. 1995) and Teufel and Moens (1997, 2002) illustrate fairly straightforward approaches to feature types that might be used for extractive summarising. However machine learning may also be applied to richer source information. Thus Marcu (2000) trained an RST-based source parser to enable him to construct discourse trees from which source nucleus clauses to form a summary could be derived, and Marcu and Echihabi (2002) were able to identify some specific discourse relations even with unsupervised learning. Leskovic et al. (2005) trained an SVM classifier on analysed sentence triples, using their linguistic features and graph relationships, to determine the source units to extract for a summary. Barzilay and Lapata (2005), on the other hand, trained to determine extracted sentence ordering for the output summary.

In these cases, machine learning has an essentially preliminary and support role in system design, though the part played by the information gained in summarising varies. It may have a dominant role so, given a learnt feature set, for example particular words or text positions, summarising is a two stage process with source feature identification followed by unit e.g. sentence scoring. Alternatively, learning may be applied, as in NLP question answering for example, to shape particular components of an overall, possibly hybrid statistical-symbolic, system.

Pushed hard, machine learning offers a rather thoroughgoing, fundamentally statistical approach to summarising that also appears to conflate the three-stage model. Thus in Banko et al. (2000), Language Modelling is applied to a training sample of sources and their sum-

maries to identify correlations that can be used to determine target summary ngrams, and their ordering, when given the ngrams in new sources. The choice of summary lexical unit, and sequencing of units in the summary, can be combined in one operation. Along the same general lines, Berger and Mittal (2000a) used FAQ data as proxy training material for extractive query-oriented summarising for which regular data is not available.

In this strategy, the training process for summarising is all-embracing: it absorbs and exploits whatever is implicit about the relation between sources and summaries and also, through this, whatever may be embodied in this relationship about the form of summaries deemed appropriate for the task purpose. This can be seen as an advantage: there is no need to analyse source properties or purposes, or output implications specifically, since whatever matters is automatically captured by the learning process. Of course the whole rests on the assumption that the training summaries are task-suited, and there is no way of investigating this beyond what is implied by the choice of source and summary units, most basically ngrams, as vehicles for correlations. Even within this framework, however, the general summarising model applies, since for new sources there will at least an interpretation stage where the units to which training-based scoring is applied have to be identifed, and a transformation stage where the results of the scoring are processed, also possibly some generation-stage tidying up. It is also the case that the generic approach can be applied to more elaborately processsed source material, as in Knight and Marcu (2002)'s work with syntactic parse data for sentence compression, illustrating another form of hybrid statistical-symbolic summarising. This is, perhaps, potential summarising since the compression is only for individual sentences (see also Turner and Charniak 2005); however Daumé and Marcu (2002) seek to compress whole texts using both sentence syntax and RST structure information.

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 all these primarily extractive methods, the second group of approaches jettisons the assumption that the basis of summarising is to reproduce (and thus essentially select) some of the source text (though of course individual lexical items may carry forward). Even approaches that may prune and merge source sentences or constituents (e.g. Newsblaster, McKeown et al. 2002), are essentially extractive. While it has become common to refer to non-extractive summaries *abstracts* rather than extracts, this tends to carry with it the implication that the goal is summaries of the informative kind conventionally labelled "abstracts" in scholarly and technical publications. However the techniques involved may be appropriate to the many other kinds of condensation represented by, e.g., a synopsis, or a review.

The methods investigated under this heading (early illustrated by DeJong 1982) have generally sought to dig below the source linguistic surface to identify conceptual content and more specifically, to determine particularly important content. As a natural corollary, source sentence analysis tends to become more ambitious, and overall discourse analysis more discriminating, though there is great variation. In general, deeper symbolic sentence analysis, giving predicate-argument, logical form, etc. sentence representations is correlated with symbolic discourse structures, with explicit discourse relationships like 'Consequence' playing a part in signalling concept status and significance in texts as wholes.

However as is evident from the discussion of extractive summarising. summarising systems may on the one hand combine statistical models of discourse with symbolic sentence processing that could, in principle, deliver eventual non-extractive summary text; and on the other can combine symbolic discourse models with extracted sentence or phrase delivery. In general, nevertheless, digging deeper below the source surface text implies symbolic processing at both sentence and text levels, with source and summary representations showing both levels of structure, and substantial transformation operations to proceed from source to summary representations. Thus while specific task applications may simplify by selecting content for just a few discourse categories or relations, discourse structure can be expected to play a significant role in non-extractive summarising. Again, the natural corollary of digging below the source surface for underlying concept representations is that the final summary text is generated de novo from the derived summary representation ones.

As in the earlier extractive case, different types of discourse structure discussed in Sparck Jones (1995) have been used for non-extractive summarising: for example domain world structure in DeJong (1982) and communicative structure in Reithinger et al. (2000). However linguistic structure based on general rather than domain-oriented relations appears not to have been used for non-extractive summarising, though approaches like Tucker and Sparck Jones (2005)'s, which build logical form representations for source sentences, could in principle be used in ths way to deliver new text. Both top-down and bottom-up versions have also figured: thus DeJong uses a top-down domain schema, Reithinger et al.'s negotiation objects illustrate simple top-down communicative structure. Hahn and Reimer (1999)'s domain relations appear more bottom-up than top-down. Hahn and Riemer do not generate output summaries from their representations but these other systems do. Again, as with extractive summarising, systems can combine more than one discourse structure type, and both symbolic and statistical structures, as in Hahn and Reimer, which uses both strong domain relations and statistical lexical 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 illustrates 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 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 robust symbolic sentence processing to logical forms is a challenge, especially for whole texts, symbolic discourse structure determination is a challenge, and relating the two to drive text condensation

is a challenge. However as Hahn and Riemer point out, deeper source representations may have the advantage that they are hospitable to a variety of subsequent summarisation uses.

Conclusion on system characteristics

It is evident that while summarising systems may be broadly categorised as extractive or non-extractive, there is enormous variation in the detail, and that systems vary widely in complexity and processing effort. They differ in particular in the treatment of discourse structure as a key guide in summarising, even if many systems make use of some kind of statistically-based salient topic identification. Where some systems use just one type of structure, often applying $tf * idf$-type scoring to determine content importance, Elhadad and McKeown (2001), for example use four structures of two types: linguistic structure to identify pertinent material in the source and domain template structure to represent this, with a different, derived domain graph structure for the summary representation and another linguistic rhetorical structure within generation to produce the text output. Many more systems use statistically-based implicit meaning representation(s) than explicit symbolic ones, but there are also hybrid approaches and some cases where symbolic structures play the major role.

In some cases the complexity follows from a rather specific application context, as in Elhadad and McKeown (2001), but in others reflects a more general summarising strategy and also one designed to raise summarising standards , as in Harabagiu and Lacatusu (2002). In most systems, source interpretation is the most complex stage, creating a source representation that sets the scene for subsequent processes such that transformation and generation are comparatively straightforward. However tranformation may be more demanding and lead to a new form of representation, as in Elhadad and McKeown, and this may in turn be subject to further reformulation, again as in Elhadad and McKeown. The same applies to Saggion and Lapalme (2002).

In volume terms, many systems use only weak forms of discourse structure as means, in particular, of interpreting source texts and presenting their content. Thus a single statistical salience structure may serve as the base for the whole summarising operation. Other systems illustrate a much richer use of multiple structures, as in Elhadad and McKeown, but there are far fewer of these.

Taking both extractive and non-extractive approaches together, the discourse structures systems have used, and the ways they have used them, have been scattered over a wide space of possibilities. It is therefore impossible 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 these choices are clearly motivated by the application task, e.g. the legal structures used in Farzinder and Lapalme (2004), and in such cases the structure's role and value seem relatively clear, at any rate in the large if not in detail. But in others the summarising context is not given, especially not in any detail, and so cannot be taken as motivating the summarising strategy. The lack of comparative evaluation, and even any evaluation at all, makes it difficult to judge strategies' relative merits. Even where there has been careful evaluation, as in DUC, it does not support strong conclusions about system strategies. So far, perhaps all that can be said in general about discourse structure in summarising, on the basis of the work done to date, is that the weak linguistic structure associated with lexical repetition does seem to be useful for determining topics and topic salience. This is not to suggest that richer discourse structures do not matter: they clearly do for language-using

tasks in general. It is rather that, recognising also that discourse structure is hard to capture, summarising research has not so far been able to make good use of such structure outside limited contexts. But it is alos possible that individual applications differ so much in their factor detail that we cannot expect much strategy portability, and have to fall back on weaker generalisation.

Factor influences on strategy choice

Thus in reviewing recent work, while some approaches follow directly from the task context, for exampke McKeown et al. (1998), much more work seems to follow either from some generic view of summarising without a detailed task and context analysis, and perhaps also without making sufficient allowance for the many forms that summaries can take; or from the very different starting point of an experiment to see whether whatever is available as current technology could suffice for task needs. This last is the normal rationale for the bulk of extractive work, though an extractive strategy may also be consciously justified as one matching task needs, for example for indicative summaries suited to literature assessment in retrieval.

But the underlying problem, in trying to assess task-strategy matching, is that evaluation tasks and performance measures so far have not generally been taxing enough to force the thoroughgoing analysis of factors that would seem in principle to be required to guide summarising system design for a specific task. This would be far from easy to do: for example with respect to input factors, what are the fine-grained features of the source genre, say, and how much do they matter, in detail? For example, do different sources require different treatments of nominal groups in interpretation and transformation? Again, while we may see different authorial styles in scientific papers and want to have source analysers capable of parsing sentences in different styles, do we need to adjust our summarising procedures explicitly so we summarise dogmatic and tentative authors differently? Even though individual author styles may have some effect on major concept recognition for summarising, they may matter less than systemic language features or, even if we did respond to them, make no real difference to the overall utility of the summary for its task purpose.

Exemplar systems

I considered broad classes of summary system primarily from the point of view of the kinds of information they manipulate and representations they use. This section considers some selected exemplar systems as wholes, with the aim of showing how varied the approaches to automatic summarising that have been developed in the last decade have been, though their performance has been much less varied. The exemplars also illustrate the somewhat eclectic character of many systems, though some have what may be called a dominating philosophy.

I have taken systems without the leverage of query orientation, or degree of reduction of headlines, and ones subject to some robustness tests, e.g. in DUC. The systems are mostly well known but my focus here is in comparisons of strategy and structure across the whole range rather than between essentially similar systems. The fact that they are mainly (though not exclusively) multi-document summarising systems reflects the somewhat fortuitous way the field has developed.

MEAD (Radev et al. 2001b, 2004)

MEAD is an essentially statistical system applicable to either single or multi-document summarising. For single documents or (given) clusters of documents it computes a term-centroid topic characterisation based on tf and idf information. It ranks candidate summary sentences using a combination of (a) sentence score against the centroid, (b) text position value (declining from source text beginning to end, and (c) $tf * idf$-style overlap with each source's title/first sentence. Sentence selection for the summary is constrained by the desired length, and by a test for redundancy given already-selected sentences based on on cosine term similarity. There are also user options, e.g. specifying length conditions for selected sentences.

The forms of representation used, for both sources and summaries, are simple term vectors and score sets, and the processing is equally simple in both interpretation and transformation as it consists only of vector comparisons and score computation. Nevertheless MEAD performed respectably, at middle system level, in DUC 2001 and 2002, is used as a summarising component of other systems, e.g. NewsInEssence (*NIE*), and is now a public domain system.

Newsblaster (NWBL, McKeown et al. 2002)

Newsblaster is a primarily statistical system but with important symbolic processing elements. It is a fully operational public system and thus includes operations, and addresses concerns, that do not figure in the mostly experimental systems described in the literature. Thus it has initial steps to identify documents that are news stories, and to cluster these, and final steps oriented towards convenient and informative online viewing, for example by adding images, as well as its main summarising procedures. Clustering is multi-level, with statistical grouping using both $tf * idf$ word information and syntactic features, and a top-level assignment to prior broad news categories. News stories are typed as about single or multiple events, or people, or 'other', with different summarising strategies invoked according to type.

For events, for example, summarising for a document cluster starts by finding similar text units (paragraphs), taken as defining themes, again using both classical statistical information along with simple and combined symbolic features. Full symbolic parsing is then applied to the sentences for a theme, and the parses are compared to identify semantically and syntactically similar components while allowing for paraphrase and other variations: these similar items are fused and, as they are taken as important because they occurred several times, are selected for the output summary. The resulting set of phrases (i.e. their representations) is then ordered by their original temporal appearance and input to a text generator which combines them where appropriate and fills them out as needed to produce complete sentences for the final summary. Summaries for other source types are somewhat different, since they use models of the sort of information that is likely to be important, within the common system framework.

Overall this is a sophisticated system which includes training to identify useful features for computing text unit similarities. It serves to emphasise the range of representation and processing options possible within the overall structural model of Figure 1. In this case the source representation seems to be no more than the sets of source text units grouped by theme and their parse trees. The main work is done in the transformation stage where the source sentences are parsed, compared and pruned, with the summary representation as the selected phrase parse trees. The summary representation is deeper than the source one, so the generation stage to deliver text is also substantial.

From the structural point of view, Newsblaster's central component performed very respectably 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 one based on a very different approach to Newsblaster's. It is designed to produce both single and multi-document summaries, and both extracts and abstracts. However the single-document strategy is rather straightforward, and it is the multi-document summarising that is of real interest. The abstracts are extraction-based: the justification for calling these multi-document summaries abstracts is that they are based on information extraction (IE) techniques.

Thus GISTexter multi-document summarising uses IE-style topic templates, either from a prior set or, if the topic is new, by adhoc template generation. After initial sentence parsing and co-reference resolution, the core system process maps source-text snippets onto template slots using pattern rules, and notes co-reference relations. For a given set of source documents templates are classified as, e.g., about the dominant event or subsidiary events: the co-reference notes make it possible both to identify main events and to provide common forms for references in the output summary. From the point of view of system generality given open-domain source, the template creation procedure is of particular interest. This exploits WordNet to identify topic relations that define semantic roles for key source lexical items. The linguistically-oriented templates that result are less powerful than hand-crafted IE ones, but are still found useful. The summary generation process using the templates is conditioned on the amount of available filler material, the template class, and the required summary length: it invokes source sentences for snippets and outputs these along with suitable reference expressions in original order, and with pruning of non-snippet material to satisfy length constraints. (The system apparently always generates well-formed sentential output, though whether this is because snippets are always clauses, or through use of the initial parsing data, is not clear).

GISTexter performed well in DUC 2002, even though it required some novel templates. It is an elaborate, resource-rich system, since it relies on templates, either existing or built via WordNet, and a set of template-construction procedures, and includes complex processes for co-reference management and for summary derivation from templates. The source representations consist of the filled, and reference-annotated, templates; given these, the tranformation-stage is mainly selective of templates, and the generation stage, as logically separable, consists of source sentence invocation and, where appropriate, pruning. GISTexter was replaced by Lite-GISTexter in DUC 2003, but this followed from the evaluation task specifications, for which the full GISTexter IE model was not appropriate (Lacatusu et al. 2003).

Verbmobil (Reithinger et al. 2000)

Reithinger at al. illustrate a very different summarising situation and approach. This is for multilingual human dialogues in a limited (travel) domain. Summarisation is thus just one function in a system primarily devoted to speech transcription and automatic translation between speakers using different languages.

Transcribed utterances are processed, using a hierarchy of parsers, to extract dialogue acts and their domain content. Content is mapped into domain topic templates, with dialogue act

operators. These act-content units are grouped into 'negotiation objects', e.g. PROPOS[AL], becoming more specific as the dialogue progresses.

Summarisation, which is user-requested, is based on the most complete accepted negotiation object for each major travel category (e.g. accommodation, travelling). Summaries are generated in, e.g., German or English, using discourse and sentence planning, with the content for each information category packaged using appropriate verb fillers etc. and discourse control, e.g. to maintain focus for multi-sentence paragraphs. Reithinger et al. report a limited evaluation of the summarising component, though Verbmobil itself was a very substantial enterprise, with its own primary evaluation for translation.

Reithinger et al. illustrate summarising in a very different environment than the usual news data ones. Their approach is specifically geared to dialogue, with its emphasis on dialogue acts, while taking advantage of a limited and well-specified domain. It is based on rich symbolic processing, with a little help from statistics for dialogue act identification, and exploits both a domain world model embodied in its templates and communicative models of dialogue and negotiation. The source representation is a set of negotiation objects, the summary representation a selected subset of these objects. The major processing effort is in input interpretation and in output generation; the transformation stage is simply an object selection one.

6 Conclusion

There is no doubt that 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, somewhat surprisingly, with open-domain sources and wide and varied conditions. Some of these systems are very simple, notably the Web search engine summarisers for retrieved document lists, but are useful nonetheless; and others, like Newsblaster, are much more sophisticated, though how useful they are, and to whom, is not clear. The same applies to, e.g., the Microsoft summariser which has been taken as a comparator in a range of tests and does not perform, in them, especially well. In some of these cases, the summariser benefits from the guidance a query gives, in others the summariser may benefit from weak user interests just in getting something indicative, but more substantial, than document titles.

Summarising research has benefitted from work on neighbouring tasks, notably question answering as well as document retrieval. It has also benefitted, quite frequently though sometimes only informally, from corpus training data. More importantly, it has benefitted from the evaluation programmes of the last ten years, most obviously DUC but also NTCIR and the related TREC QA programme. These evaluation activities have been important both for their direct contribution to the development of evaluation methodologies themselves - even if there is still far to go, from the results obtained in the successive evaluation rounds, and from the way they have encouraged researchers everywhere to address the issues of task specification and performance assessment.

More specifically, 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, extractive summarising without support from domain grammars or ontologies and relying on statistical methods, perhaps with shallow linguistic processing as well. There is some evidence, though real world data is very patchy (i.e. little more than anecdotal), that these techniques can be operationally useful where crude or minimal summaries, or phrasal

summaries, are sufficient for purpose. It is in particular possible, though again substantive real-world data to support this view are lacking, that summarising that combines staistical with light-weight symbolic language processing can be more useful than purely statistical methods.

There is no reason to suppose, moreover, that while summarising research has indubitably benefited from being fashionable, it will not continue and seek to address the harder issues that need to be tackled.

But offsetting these positive advances, the work and evaluations done have been limited and miscellaneous when considered in the overall summarising space as, for example, discussed at the Dagstuhl Seminar in 1993 (Endres-Niggemeyer et al. 1995). The work on extractive summarising has been picking the low-hanging fruit, and the overall trend has been more technological than fundamental. There has been little work on deep approaches that build content representations far from the surface source text, that address summarising as condensation involving content generalisation as well as content selection, that engage with purposes that require radical transformation of the content and expression of the source, or that fully exploit the structures of discourse. It is not that any of these have been shown to be irrelevant: it is much more, as Marcu (1999, 2000)'s experiments with Rhetorical Structure Theory showed in relation to structural analysis for example, that we do not know how to automate these challenging processes. Thus except for specific applications, we do not know, for example, how to identify source content structures richer than those associated with lexical repetition; or how, except where non-linguistic sources force some provision of summary structure, to replace source structures by new, purpose-specific ones.

As a natural corollary, particularly when combined with the difficulty of characterising the tasks for which summarising is intended and of evaluating performance for these tasks, we cannot say much about the types of operation or representation they require, or at least with which they might be better served than those so far tried. There is a lesson here in the TREC programme. This began with a rather 'conventional' view of the retrieval task, as established by preceding research. But over time TREC branched out with more detail per task and more tasks. We should see DUC and its sister programmes as beginning to seek, but so far only in a modest way, a better understanding of summarising and ability to automate it. There is no reason to suppose that while summarising research has indubitably benefited from being fashionable, it will not continue and seek to address the harder issues that need to be tackled. Thus we should drive this research with more challenging formulations of the task through a wider and more demanding range of factor, especially purpose factor, specifications; and, accompanying this, through a finer treatment of the intrinsic/extrinsic evaluation range.

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.

Appelt, D.E. et al. (1993) 'FASTUS: a finite-state processor for information extraction from real-world text', *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence (IJCAI-93)*, 1993, 1172-1178.

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.

Chen, F.R and Bloomberg, D.S. (1998) 'Summarisation of imaged documents without OCR', *Computer Vision and Image Understanding*, 70 (3), 1998, 307-320.

Chinchor, N.A. (1998) 'Overview of MUC-7/MET-2)', *Message Understanding Conference Proceedings, MUC-7*, http://www-nlpir.nist.gov/related_projects/muc/muc_7_proceedings/ (visited April 2006).

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.

Kintsch, W. and van Dijk, T.A. (1983) *Strategies of discourse comprehension*, New York: Academic Press, 1983.

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.

Li, Y., Zhang, T. and Tretter, D. (2001) *An overview of video abstraction techniques*, Report HPL-2001-191, HP Laboratories, Palo Alto, 2001.

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., Bloedorn, E. and Gates, B. (1998) 'Using cohesion and coherence models for text summarisation', *AAAI-98*, 1998, 69-76.

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', *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. (1999) 'The automatic construction of large-scale corpora for summarising research', *Proceedings of the 22nd Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1999)*, 1999, 137-144. (1999b)

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 an 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 sumariser 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', *DUC 2005*, 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: Bungley: 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.

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.

Toklu, C., Iiou, A.P. and Das, M. (2000) 'Videoabstract: a hybrid approach to generate semantically meaningful video summaries', *International Conference on Multimedia and Expo (III) (ICME2000)*, 2000, 1333-1336.

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.

Turner, J. and Charniak, E. (2005) 'Supervised and unsupervised learning for sentence compression', *ACL 2002: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, 2005, 290-297.

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.

White, R.W., Jose, J.M. and Ruthven. I. (2003) 'A task-oriented study on the influencing effects of query-biased summarisation in web searching', *Information Processing and Management*, 39, 2003, 707-733.

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.

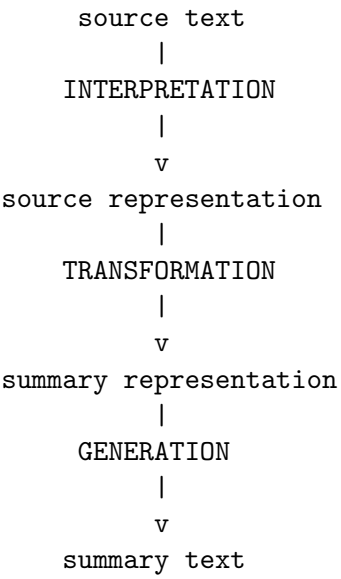

Figure 1: Schematic summary processing model for text

```
  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