Factorial summary evaluation

Karen Spärck Jones
Computer Laboratory, University of Cambridge
William Gates Building, JJ Thomson Avenue, Cambridge CB3 0FD, England

Invited talk at the Workshop on Text Summarisation, ACM SIGIR 2001 Conference, 13-14 September 2001.
www-nlpir.nist.gov/projects/duc

In this note I will develop the theme introduced in Sparck Jones (1999), namely what the *context factors* are that affect summarising and what they imply for evaluation, illustrating my points with detailed examples. My claims are first, that there are many varieties of summary and this variety has not been addressed in work on automatic summarising. Second, that this variety arises because the contexts in which summaries are used differ and differ in more ways than is generally recognised. Third, that these context factors have a large impact on summarising. Fourth, that significant implications follow for summary evaluation, since for proper evaluation it is necessary to apply context, and in particular the *task* for which summaries are intended.

I am deliberately taking a broad view, and therefore start from the definition that *A summary is a reductive transformation of source text to summary text through content condensation by selection and/or generalisation on what is important in the source.*

1 Summary varieties

In general, research on automatic summarising so far has not analysed context factors - i.e. *input*, *purpose* and *output* factors - and their implications in depth. This is partly for the good reason that getting any summarising technology up and running is so hard it is sufficient for initial assessment to apply very basic evaluation strategies. These have primarily been *intrinsic*, focusing on system success in capturing key source concepts and in producing comprehensible summary texts. However the difficulty of summarising has meant that the distinction between *informative* and *indicative* summaries has been recognised, particularly since indicative summarising seems easier and may be adequate for some undemanding purposes, like assessing potential source relevance to information needs. But such purposes have still been treated in a rather general way, with specific evaluation data (e.g. search topic sets) treated as typical for a range of contexts.

Indeed, work on summarising has usually taken it for granted that the aim is to produce *generic*, i.e. multi-purpose, summaries. Generic summaries are often regarded as neutral,

when they are in fact reflective of their sources, e.g. in language and hence envisaged readership. When existing, independent human summaries (i.e. not ones specially constructed for evaluation purposes) are used for comparative evaluation there is a similar, but again frequently unrecognised, implicit reference to context factors, for instance in the retention of source technical terms in a scientific abstract.

Some research, for instance McKeown et al's work (1998) on multidocument summarising in the medical domain or Schiffman et al's (2001) on biographical summaries for analysts, is explicitly geared to a particular application context, and may have been evaluated in that context. However even here the evaluation may be intrinsic, assessing the extent to which the system delivers the required sort of summary, rather than *extrinsic*, actually addressing functional fitness in use. Moreover generic summaries are most commonly contrasted with *query-based* summaries, i.e. ones delivered in response to a search, though even here the focus is on summaries that relate to the topic of the search alone, without taking any other aspects of the search requirement e.g. for summaries drawn from documents of a particular type.

2 Factor multiplicities

As the foregoing suggests, there has so far been little indepth analysis of context influences on summarising. This is unsatisfactory for two reasons. First, the number and complexity of context factors is far greater than is usually recognised. Second, they have large consequences for evaluation, even if we allow that there may be many equally good summaries for given sources because it is impossible to specify context requirements to closely as to rule out, for instance, many variant output texts. In general the condensation of source to summary is so drastic that, given both are complex linguistic objects, there are many ways of doing it. This does not, however, imply that there is no point in trying to constrain summarising. Context factors, especially the purpose(s) for which summaries are required, are so important it is essential to try to match them as well as possible.

At the same time, automatic summarising has potential advantages compared with human summarising. It should in principle be much more open to context-sensitive processing, and allow far more for alternative summaries, because the costs with established systems can be expected to be much lower than for human summaries. This requires, however, not only far more powerful, i.e. flexible, systems than any so far. It depends on being able to specify context requirements well enough to set system parameters appropriately.

Figure 1 shows a short source text to be used for illustration in what follows, an imaginary information report about wombats put out by a police force (with, it seems, somewhat literary aspirations). Figures 2,3 and 4 provide input, purpose and output characterisations for a particular envisaged summary application (Purpose 1), which we suppose is providing warning alerts about traffic hazards for a local newspaper. The factor lists substantially extend those of Sparck Jones (1999), and also reflect a more careful analysis of output factors. Purpose factors are the dominant ones, since a summarising system should be designed to satisfy these. However it is clearly necessary, not only in order to build an automatic input processor but to respond appropriately to purpose factors when summarising, to take input factors, i.e. the properties of the source, into account. Output factor choices are then made to deliver summaries that satisfy the constraints imposed by input and purpose, illustrated by the final summary shown in Figure 4.

In this case, given that the example source is simply a descriptive report but the summary

is intended to alert, it is not sufficient to extract (whether at the surface or in some more sophisticated way) a statement to the effect that some wombats have been found on the ground: a reworking of both source information content and linguistic expression is required. Thus the purpose has implications both for the nature of the summary and its relation to the source, since the appropriate focus is on issuing a warning about something unusual to attend to, and without regard to the reasons why this is occurring. The input factor characterisation of Figure 2 may seem too obvious, but as both the point just made and the contrast sources given in the figure imply, an explicit, detailed chacterisation of source properties is required for effective summarising. At the same time, though the purpose specification imposes strong constraints on the output summary, these are not absolute: thus alternative wordings, e.g. `'Watch out for ...'` or `'... on the road'` or `'Wombats by the road: watch out!'` could all satisfy them.

3 Factor effects

Doing a detailed decomposition of factor effects for different applications shows how summaries with quite distinct characteristics can be derived from the same source document (or set of documents). Figure 5 illustrates an alternative purpose requirement - supplying briefs about wombats to a central scientific information site - and a summary that could naturally follow from this requirement along with an associated output factor specification.

This second summary is quite different from the first, and is fundamentally, not merely superficially, so, because the two purposes they are designed to serve are quite distinct. The difference between the requirements is, moreover, too great to be effectively overcome by a single *hospitable* multi-purpose summary. The information about wombat eating behaviour that is important for Purpose 2 would be a distraction in a summary for Purpose 1. The instructional genre suited to Purpose 1 would be be unsatisfactory, on the other hand, for Purpose 2. It is true that a single minimal summary `'Wombats on ground!'` might not be without some utility in both cases, but would be far from ideal for Purpose 2. Of course in individual cases, purposes may be only inadequately or broadly known, or several have to be served simultaneously, so more tolerant, compromise summaries are in order. But these will normally be less than optimal for any individual context.

At the same time, though purpose factors are the key constraints on output summaries, and may be especially strong for individual, known applications, it does not follow that they dictate even material properties, let alone every detail, of the output specification. Purposes can usually only be characterised in such general terms as 'audience: research biologists'. Even when summaries are intended for a known individual (as with an executive summary for President Bigman's decision on pineapple subsidies), there can be variation in substance and form. For the illustrative example, Purpose 1 as specified could lead either to the output factor specification 1.1 and associated summary already given in Figure 4, or to the rather different one, 1.2, that combines text and graphics, shown in Figure 6. It is not clear whether the resulting alternative summary is better or worse, in itself, than the earlier one for the purpose in hand.

These illustrations emphasise the fact that summarising is a source-to-summary condensing *transformation*. Thus for the alerting summary 1.1, for instance, there are large changes in relation in content - the summary is selective for local daily life and omits material on wombat diet; in presentation - the summary is attention grabbing, not a descriptive account;

in language - the summary is popular informal in style, not regular prose; and in reduction - the summary compression is drastic and absolute (not by percentage). This summary is certainly not simply text or concept 'extractive' The same applies, with the introduction of the image especially, to 1.2.

Summary 1.1 might be deemed extreme as a summary, perhaps not a summary at all, because producing it requires not only some rather sophisticated discourse processing and inference, but also the introduction of the task-motivated concepts represented by the injunction `'Keep an eye open for'`. However, even if we were to allow the claim that 1.1 is not 'really' a summary, rather a (new) warning, alternative outputs that might be both accepted as summaries and judged suited to the alerting purpose can require similar deep source processing. Figure 7 illustrates a descriptive statement as an alternative to the original 1.1, and an analogous alternative to 1.2. These are certainly summaries but nevertheless similarly radical transformations of the source.

4 Evaluation implications

As mentioned, there are good reasons for beginning evaluation with intrinsic methods. It is more difficult to rely on notions of correctness than with machine translation, for example, especially for generic summarising that is intended to meet many needs. But it may be reasonable to work with looser notions of plausibility, appropriateness and so forth, especially where context requirements are fairly clear so whether summaries may meet them is not too difficult to judge. However the range of alternatives that can be produced to meet even quite detailed and careful purpose specifications suggests that assessing summaries by their likely utility is a rather weak evaluation strategy. It is also possible to ramp up intrinsic evaluations to obtain finer numeric performance metrics, by developing and applying source and summary markup schemes for key concepts. But it is then essential to recognise that the markup is either making implicit reference to context requirements that would be better made explicit - which includes acting on the presumption that summaries should simply *reflect* source information and expression; or is based on the not necessarily sound assumption that human judgement - even agreed judgement - on key source elements is an adequate way of defining the proper content for a summary regardless of potential use.

Thus even where the context requirements are clear, so judgements on summary acceptability seem unproblematic, it will always be preferable to test summary utility for purpose directly. The examples in the previous section make it quite clear both that different contexts demand very different summaries, and that competing alternatives for a given application may be hard to assess apart from actual usage.

But we then find, first, that particular task requirements may be rather undemanding, as with indicative summaries for document retrieval, so performance 'extrapolation' to other tasks is rash; and second, that task-based evaluation is very difficult and, typically, very expensive.

This evaluation challenge is illustrated by the scenario shown in Figures 8-10. This scenario is for a task-based evaluation for alerting summaries like those of Purpose 1, working from the particular starting point that prompts the evaluation and envisages a fairly modest initial assessment of a new alerting process. This process exploits as (assumed) automatic system for producing summaries, but the evaluation is not of the system alone. The evaluation is of the operational *setup* in which summaries are used, i.e. is an evaluation which is extrinsic

to the summarising system but intrinsic to the setup as a whole. The scenario follows the model presented in Sparck Jones and Galliers (1996), establishing the evaluation remit, and the detailed design, dealing with performance factors and performance assessment, evaluation data and evaluation procedure. The point of the decomposition, especially in the remit and characterisation of the evaluation 'subject', is to motivate the particular design chosen. The evaluation is naturally for more than one wombat alerting occasion: I assume summarising is from a variety of sources about wombats variously invading roads and constituting potential traffic hazards, for instance young wombats using roads as playgrounds in spring.

Even though the evaluation is only sketched, and is no more than a modest indicative investigation, the scenario makes it clear that a properly-conducted task-based evaluation is far from trivial. Characteristic issues of detail are illustrated in Figure 11 showing, for example, how difficult it would to formulate a question set to gather useful and reliable information about the newspaper alerts for the questionnaire-based evaluation envisaged. If the alerting system was not restricted to wombat hazards, this would be much more tricky. Further, different motivations and hence remits would lead to different scenarios, for instance for a comparative evaluation designed to establish the relative utilities of the two styles of output, text or hybrid, for Purpose 1. Another, extrinsic rather than intrinsic, evaluation would be required to discover whether the alerting setup was helpful to the local traffic police (as opposed e.g. to sending round loudspeaker vans). More importantly, as indicated in Figure 12, it is clear that task-based evaluation for Purpose 2, designed to assess the value of the summaries contributed to the Wombat Information Web Site, would have to take a quite different form, though it would present analogous problems of design and implementation.

Such task-based evaluations are necessarily specific to individual contexts. But well conducted ones can be used, just as with document retrieval, to establish relative performance for different systems, and general levels of performance for data conditions, in a fairly reliable way.

Conducting even one, let alone several, task-based evaluations may appear a daunting prospect for summarisation research. The foregoing should not be taken to imply, either, that intrinsic *system*, or even subsystem, evaluations have no useful role. But they have to be firmly context-related. Thus just because there can be no objectively correct summary of a source, it is essential to make the contextual assumptions that underlie 'natural' or 'consensual' reference data, whether human summaries or source analyses for key concepts, quite explicit. All the context factors bearing on summarising have to be laid out in detail, especially through the system or setup's *environment variables*, in order to characterise the sort of summary that is required sufficiently well to allow some helpful, even though only indicative, initial evaluation. With explicit statements of the task assumptions on which summarising is based, we can both get better-grounded intrinsic evaluations and make a start on the long, necessary haul to task-based evaluation.

References

McKeown, K., Jordan, D. and Hatzivassiloglou, V. 'Generating patient-specific summaries of online literature', *Intelligent text summarisation*, Working Notes, AAAI Spring Symposium 1998; Menlo Park, CA: American Association for Artificial Intelligence, 1998, 34-43.

Schiffman, B., Mani, I. and Conception, K.J. 'Producing biographical summaries: combining linguistic knowledge with corpus statistics', *Proceedings of the 39th Annual Meeting*

of the Association for Computational Linguistics and the 10th Conference of the European Chapter of the Association for Computational Linguistics, 2001, 450-457.

Spärck Jones, K. 'Automatic summarising: factors and directions', in *Advances in automatic text summarisation*, (Ed. I. Mani and M. Maybury), Cambridge, MA: MIT Press, 1999.

Spärck Jones, K. and Galliers, J.R. *Evaluation natural language processing systems*, Lecture Notes in Artificial Intelligence 1083, Berlin: Springer, 1996.

```
Source text :
   Northtown Police report. January 3, 2004
   Recent reports from local farmers suggest that the wombat
   population is rapidly increasing, since the wet weather has
   provided unusually lush vegetation for them to eat and plenty
   of cover to sleep in undisturbed. In some cases, however,
   it seems that the wombats have eaten so much that they are
   unable to maintain their grip on tree trunks while sleeping
   and, sliding down, have been discovered in the early morning
   by passing farm workers, slumped gently snoring on the ground.
```

Figure 1: Example source

```
Input factors :
  Form :    language - educated English
            medium - text
            structure - header, running text
            genre - descriptive report
            length - 85 words

  Subject type : class - ordinary
                 level - lower
  Unit : single
  Author : one
  Header : originator, date

Contrast sources :
 'Fossil wombats' archaeological paper
 'Wendy the wombat' children's tale
```

Figure 2: Input factors

```
Purpose factors 1 :
   Use            - alert
   Audience       - local population (large, varied so only semi-targetted)
   Envelope
     time         - 24 hour currency
     location     - Northtown News
     formality    - none
     trigger      - report receipt
     destination  - direct user consumption

(Summary : emphasis on unusual presence of somnolent wombats

   e.g. Keep an eye open for wombats right by the road.      )
```

Figure 3: Purpose factors, Example 1

```
(Purpose factors 1 : alert, locals, 24 hours, News)

Output factors 1.1 :
  Material
    coverage      - bearing on town population's daily activities
    reduction     - to single sentence
    derivation    - no restriction
    speciality    - everyone local
  Style           - informative core
  Format
    language      - popular English
    medium        - written
    structure     - no special layout
    genre         - instructional

(Summary :
    Keep an eye open for wombats right by the road.      )
```

Figure 4: Output factors, Example 1.1

```
Purpose factors 2 :
   Use            - substitute primarily, also preview, filter
   Audience       - international wombat researchers
   Envelope
     time         - open
     location     - Wombat Information Web Site
     formality    - none
     trigger      - none
     destination  - summaries may be further processed e.g. grouped
                     by theme, for new summary syntheses

(Summary : note fact of and reason for wombat behaviour in
            specific area

   e.g. Unusually rich vegetation in Northtown, January 2004,
        encouraged wombat population growth and gross overeating,
        so wombats slept on the ground.                          )
```

Figure 5: Purpose factors, Example 2

```
Output factors 1.2 :
  Material ...
    reduction     - less than 2 inches of column space
  Style ...
  Format ...
    medium        - hybrid, drawing and text
    structure     - headed text

(Summary :
    ATTENTION: WOMBATS !
        mmmmmm
      mmmmmmmmmmm          < i.e. silhouette of slumped wombat >
    sleeping along the roads   )
```

Figure 6: Output factors, Example 1.2

```
Text alternative to summary 1.1

‘Wombats unexpectedly sleeping around’


Text + graphics alternative to summary 1.2

        mmmmmm
      mmmmmmmmmmm  <silhouette>
    sleeping where you walk
```

Figure 7: Variations on Examples 1.1 and 1.2

```
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
```

Figure 8: Evaluation remit, Example 1.1

```
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
```

Figure 9: Evaluation design, Example 1.1

```
 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
```

Figure 10: Evaluation data and procedure, Example 1.1

```
Issues - the devil's in the details :

 population sampling -
 random for each of publication day
                    second (repeat) day
                    three days later (late reading/memory)

 question design -
   Q1 What do you know about wombats round here ?
   Q2 Have you seen any wombats on/by the road ?
   Q3 Have you hit any wombats ?
   Q4 Did you see anything about wombats in the paper ?
      if yes
      Q4a What did it say ?
```

Figure 11: Detail problems in evaluation, Example 1.1

```
Evaluation variants :
 alerting - intrinsic, text better than hybrid
 alerting - extrinsic to setup in assisting traffic police


Evaluation for alternative purpose (Wombat Information) :
 Goal - establish summaries enough information for researchers
 ...
 Design - determine use of Web Site as source of data for
           papers on wombats
 ...
```

Figure 12: Evaluation variants, Examples 1.1 and 1.2, Example 2