On the subjectivity of human-authored summaries

BALAKRISHNA KOLLURU and YOSHIHIKO GOTOH

Department of Computer Science, University of Sheffield, Sheffield S1 4DP, United Kingdom
e-mail: {b.kolluru, y.gotoh}@dcs.shef.ac.uk

(Received 14 May 2007; revised 18 February 2008; accepted 24 April 2008; first published online 2 June 2008)

Abstract

Human-generated summaries are a blend of content and style, bound by the task restrictions, but are ‘subject to subjectiveness’ of the individuals summarising the documents. We study the impact of various facets that cause subjectivity such as brevity, information content and information coverage on human-authored summaries. The scale of subjectivity is quantitatively measured among various summaries using a question–answer-based cross-comprehension test. The test evaluates summaries for meaning rather than exact words based on questions, framed by the summary authors, derived from the summary. The number of questions that cannot be answered after reading the candidate summary reflects its subjectivity. The qualitative analysis of the outcome of the cross-comprehension test shows the relationship between the length of a summary, information content and nature of questions framed by the summary author.

1 Introduction

Summarisation invariably is an act of reducing a document to a fraction of its current size, yet retaining its most important message. This reduction involves removing the unwanted or redundant information from the document. It is here that the subjectivity makes its distinct presence as quite often the ideas of relevance, importance and redundancy are relative. Thus, one can possibly concur that there exist many versions of a good summary, each of which is accurate from the dimension of its author. Human beings do not totally agree on what should be a ‘one good summary’ (Mani 2001). Probably this can be attributed to the different perspective by every individual whilst writing a summary. Guided by various factors such as educational background, profession, personal interests and experience, an individual decides whether a certain aspect is worth being included in a summary. What might seem relevant to one person could be deemed redundant by another when reading the same story, thus accounting for more than one ‘correct’ summary. The issue of subjectivity gains prominence as the compression ratio increases, i.e., the shorter the

*This work was funded by UK EPSRC grant GR/R42405, Statistical Summarisation of Spoken Language (S3L).
summary, the larger the number of ‘correct’ summaries (Lin and Hovy 2003). This is possibly due to assimilation of seemingly important contents taking priority while discarding the redundant information. This relation between subjectivity and the variations of a summary indicates the need to quantify the effect of subjectivity on issues such as summary evaluation.

Although the subjectivity reflects individual’s thoughts, there will also be some information commonly observed in different summaries of the same story. Stated otherwise, words in a summary may vary, phrases may vary and often the grammatical structure may not be the same, but a certain degree of information may be common across summaries. To what degree is information uniform across different summaries? How much subjectivity is there? How do we account for similar information stated using different words, expressions or grammatical structure when comparing summaries? How does this help when gauging the informativeness? Does the subjectivity cause any adverse effects when evaluating summaries? It is these questions that we aim to address in this paper.

Let us assume that the atomic facts of a summary account for its relevance. Then, a simple question that elicits any one of these atomic facts represents a benchmark for assessing its informativeness. We wish to evaluate the quality of a summary in terms of atomic facts commonly observed in, or subjectively discarded from, assorted human-authored summaries. In our quest to quantify the subjectivity, we devise a cross-comprehension test along the lines of Hirschmann, Burger, Palmer and Robinson (1999) for extracting atomic contents. The comprehension test is modelled on a question–answer style framework. ‘Crossing’ the model turns out to be an effective scheme for measuring the divergence among multiple summaries. Questions are prepared by the subject who wrote the original summary (Section 2). Their answers should be derived by reading the summary alone. Summary-questionnaire pairs are then swapped in such a way that any summary is paired with questions written by other subjects (Section 3). The number of questions that cannot be answered by reading the summary accounts for the subjectiveness of the author (Section 4). In order to assess the effect of summary size on subjectivity, short (approximately 15 words) and intermediate length (50 words) summaries are studied for the same set of broadcast news stories. Independent judges make a decision for the relevance level of summaries given a question set; their consistency is also examined.

1.1 Previous works on subjectivity

Previous studies in information science have identified so-called general abstracting principles (GAP) which humans tend to follow when writing summaries or abstracts for documents. Cremmins (1982) established the GAP as extracting, organising, reducing and refining of information. Boret and Peyrot (1988) and Rowley (1988) also found human adherence to similar principles but more on the conceptual level, such as faithfulness to the original document, precision to use accurate terminology and entropy (conveying fullest meaning in the least number of words). The finesse shown by professional summary writers involves a complex task of information
On the subjectivity of human-authored summaries

management: abridging, refining and re-synthesising data. At the higher levels, principles of abstraction may seem to be the same; however at the lower levels, the information manipulation can be quite subjective.

Kintsch and van Dijk (1978) presented the text comprehension as a theoretical macro-structure using which the information is reduced into a gist. This macro-structure is a cognitive model based on the goals of the reader and thus affected by individual subjectivity of that respective reader. Again, the issue of subjectivity raised when text reduction was the goal. Along similar lines, Endres-Niggemeyer and Neugebauer (1998) provided an insight into the cognitive activity of summarisation. This empirical investigation laid the foundations for the simulation of summarisation as a cognitive process. Modelling a complex task such as summarisation requires careful parameterisation of the desired attributes such that the outcome of the models suits the intended task (e.g., generating indicative or informative summaries, yielding a headline or a one-line summary). Pinto Molina (1995) proposed a four-step GAP approach that would help model the aspect of document abstracting. In this study, it was proposed that reading-understanding, selection, interpretation and synthesis would be the principal attributes of abstracting.

In all these works, human interpretation of selecting a precise, succinct and coherent extract was a main underlying attribute. This, however, is a very subjective trait and is dependent on the person who is writing a summary. In other words, it is quite possible to generate, from the same document, coherent and accurate summaries that are very different from each other (Nenkova and Passonneau 2004; Hovy, Lin, Zhou and Fukumoto 2006). The counter argument for this inference is that in all the summaries pertaining to a certain document there must be some common information in spite of different expressions and styles.

Analysing a pool of human-authored summaries to gain an insight into similarity of information has been an active hot-spot in the recent summarisation research (Van Halteren and Teufel 2003; Nenkova and Passonneau 2004). Jing (2002) automated the decomposition of human-authored summaries using hidden Markov models for various applications. It was reported that initially some information was extracted from the document and then manipulated to produce a desired summary. The manipulation could include re-writing, re-ordering and generalising.

Upon further examination, Vanderwende, Banko and Menezes (2004) observed that humans tended to extract complete phrases or sentences from single documents. This, however, changed to a word bigram selection for multi-document stories. In one study concerning information coverage in human-authored summaries, Radev, Jing, Sty’s and Tam (2004) proposed to model both inter-judge agreement and system evaluation; utility scores were assigned to each candidate sentence and the relative utility was calculated for each system.

Subjectivity is also a considerable impediment to question-answering evaluation. There have been several attempts to rate answers (atomic facts or ‘nuggets’) in such a way that the scores reflect the significance of the respective answers (Lin and Demner-Fushman 2006). Dang and Lin (2007) discussed a method to combine the different scores of ‘nugget’ relevance ranking by various judges. They concurred an increased stability and a better discriminative power during evaluation.
This paper is related to the above studies in various aspects. It is concerned with the human subjectivity when they produce summaries. In particular, we propose one approach to measure the amount of subjectivity quantitatively. Although this work targets a restricted task (generic summaries with short to intermediate length) in a restricted domain (news broadcasts), we consider the problem is of fundamental importance and the idea is widely applicable to other tasks and domains. We aim to analyse human-authored summaries, associating the subjectivity with their unique interpretation of stories.

2 Production of human-authored summaries

The objective of this paper is to quantify the aspect of human subjectivity while authoring summaries. To this end, five subjects (1, 2, 3, 4 and 5) produced a short summary (approximately 15 words) and an intermediate length summary (50 words) for broadcast news stories given a simple instruction set.

Subject Profile. English is the first language for four out of five summary authors. English is the second language for one author, though very fluent because of necessity. All subjects were educated to at least graduate level. It was expected that they could produce summaries of good quality without detailed instruction or further training. A simple instruction set (discussed later) was given, leaving wide room for interpretation about what might be included in the summary. Hence subjectivity was promoted.

Data. The experiments presented in this paper were concerned with a subset of broadcast news stories from the TDT-2 corpus (Cieri, Graff, Liberman, Martey and Strassel 1999). They were used for NIST TDT evaluations and the TREC-8 and TREC-9 spoken document retrieval (SDR) evaluations. The subset consisted of 43 hours of speech; each program contained seven to eight long news stories on average, spanning 30 minutes as broadcast, which was reduced to 22 minutes once advert breaks were removed. All works in this paper were conducted using 51 stories, randomly selected from a pool of 542 news stories.1 These 542 stories were manually selected from closed caption hand transcripts, being sufficiently long.

2.1 Instructions

Summary production. A simple instruction was given to the subjects in order to craft a summary:

- A short summary and an intermediate length summary should contain roughly 15 and 50 words respectively, possibly in the subject’s own words.

1 The rest (491 stories) were kept for a different task, not related to the aim of this paper.
As the news stories ranged from 16 to 84 sentences, with the average length of 400 words in 25 sentences, subjects would have to prioritise information that could be included in their summaries. The instruction implicitly encouraged the subjects to put as much important information as possible into a summary, while maintaining a good level of fluency. It was also a flexible instruction so that the subjects were able to use their own expressions when necessary. After completion of the task, they commented that this instruction made them experiment with different words to shorten or expand the information they wanted to include. For example, how could an earthquake disaster be expressed in different ways:

8000+ feared dead? … or
thousands of people killed? … or
a lot of people are believed to be dead?

Another feature of this instruction was the amount of generalisation that a subject would like to use. For example, a subject could say

US Senate to decide on tobacco bill;

or say

Senate to vote on bill, hiking tobacco price;

while adding extra information, but omitting specific details.

**Questionnaire production.** When producing summaries, subjects were aware that they also had to prepare questions and their model answers. The following instructions were given:

- A questionnaire for a short summary or an intermediate length summary may consist of 2–4 or 6–8 questions respectively;
- An answer must be found in the particular summary, without reading the entire story;
- Yes/No questions should be avoided wherever possible;
- The summary may roughly be reconstructed from the question–answer set.

Each fact might be questioned in such a way that the particular summary could be recovered. Ideally we would expect each question to elicit a precise information point chosen for the summary – e.g., who did it, when did it happen, what was the cause? The question–answer set enabled us to gauge the most relevant information as decided by the subjects, so that their subjectiveness became apparent.

### 2.2 Summaries and questions created – a quick glance

A short and an intermediate length summary-questionnaire pairs were produced for 51 broadcast news stories by five subjects.

**Basic statistics.** Table 1 shows the average lengths of summaries for each subject. Interestingly, subjects tended to use longer words for short summaries (5.8/5.6/5.7/5.6/4.9 characters per word respectively, for Subjects 1, 2, 3, 4 and 5)
Table 1. The average lengths of summaries (in terms of the number of words and characters), and the average number of questions per summary (white spaces and punctuation are included in the character counts).

| Subject | Short summaries | Intermediate length summaries |
|---------|----------------|-------------------------------|
|         | #Words | #Characters | #Questions | #Words | #Characters | #Questions |
| 1       | 16.0   | 110.6      | 3.7        | 51.7   | 334.2      | 7.0        |
| 2       | 14.5   | 98.1       | 3.7        | 49.5   | 311.3      | 7.3        |
| 3       | 16.1   | 110.1      | 3.4        | 50.6   | 321.3      | 6.0        |
| 4       | 11.4   | 77.2       | 2.4        | 51.8   | 332.6      | 6.4        |
| 5       | 15.5   | 92.5       | 3.0        | 48.9   | 288.5      | 6.0        |

than for intermediate length summaries (5.4/5.2/5.3/5.3/4.8 characters). It may have been a consequence of efforts made by summary authors who tried to incorporate as much information as possible into a small space; when a summary was short they had to limit the contents by removing words and/or replacing a phrase with a single word. Function words, which typically are very short, are more likely candidates to be eliminated. The table also shows how the average number of questions varies between subjects.

General observation. A typical short summary produced by Subjects 1, 3 and 5 indicates a moderate level of abstraction, by using words and phrases from the original story and blending them with one’s own vocabulary. On the other hand, Subjects 2 and 4 have created short summaries with a much higher level of abstraction. Instead of using precise words in the story, they expressed the significant facts in their own words to produce a fluent and coherent summary. For short stories, most questions generated by Subjects 3 and 5 pin-pointed the exact information extracted from the story. Subjects 1 and 2 produced a spectrum of questions ranging from very precise (e.g., ‘who is dead?’) to rather vague aspects (e.g., ‘why does it differ?’) in the summary. Finally, Subject 4’s question set contained many vague and open-ended questions.

Intermediate length summaries are at least three times longer than short summaries and thus are able to carry much more factual details. For most of the summaries, the level of abstraction was not too high. Interestingly, not only Subjects 1, 3 and 5 but Subjects 2 and 4 also produced many intermediate length summaries with a moderate level of abstraction. It was observed among five subjects that most questions focussed on facts extracted from the story.

The length factor does seem to have a certain effect on the summarisation strategy. When a summary is shorter one needs to be prudent about the contents and the structure of the summary. As a consequence, a greater degree of subjectivity is unavoidable and some authors (Subjects 2 and 4) opted for a higher level of abstraction. The above aspects will be statistically examined in the experiments (Section 4).
Fig. 1. Short summary-questionnaire pairs by Subjects 1 and 2 – the answer sought by the author is given in ( ). The original is a factual story in the TDT–2 corpus, with a story ID: '19980519_1830_1900_ABC_WNT.1186'.

Short summary by Subject 1:
Lack of resources stops FDA from properly monitoring market-approved drugs, some of which cause 100000 deaths annually.

Questions by Subject 1:
1. Which organisation is being discussed? (FDA)
2. What is hampering this organisation’s efforts? (lack of resources)
3. What is this organisation trying to do in this instance? (monitor market-approved drugs)
4. How many deaths are mentioned? (100000)
5. Over what time span do these deaths occur? (every year)

Short summary by Subject 2:
Unstructured FDA is unable to perform well and in need of efficient administration.

Questions by Subject 2:
1. What is the problem? (FDA is unstructured)
2. What is the result? (unable to perform well)
3. What is the remedy? (need efficient administration)

Fig. 2. Short summary-questionnaire pairs by Subjects 1 and 4 – the answer sought by the author is given in ( ). The original is a non-factual (opinion) story in the TDT-2 corpus, with a story ID: ‘19980213_1830_1900_ABC_WNT.1103’.

Short summary by Subject 1:
Fisher’s study claims we seek partners using unconscious love maps; women prefer status, men go for physical traits.

Questions by Subject 1:
1. Who is the author of this study? (Fisher)
2. What claim does the researcher make concerning our method for seeking a sexual partner? (we use unconscious love maps)
3. What do women look for in men? (they look for status)
4. What do men go for in women? (physical traits)

Short summary by Subject 4:
Culture and chemistry both play a role in the science of romance.

Questions by Subject 4:
1. What is being discussed? (science of romance)
2. What are the factors affecting the particular event? (culture and chemistry)

Figures 1 and 2 show typical examples produced by subjects. Due to space constraints, two, out of five, pairs of only ‘short’ summary-questionnaire are presented.2

Samples. Figure 1 shows pairs by Subjects 1 and 2 for a factual news story. Factual stories are more common among news broadcasts, presenting facts from recent events. Both subjects described the cause of the problem, although they did not appear to be the same (‘lack of resources’ by the former and ‘unstructured FDA’ by the latter). Further, Subject 1 wrote the consequence, using the specific number (‘100,000 deaths annually’), while Subject 2 focused on the prescription (‘efficient administration’). Word usage among the subjects was also interesting – e.g., ‘FDA is unable to perform well’ as against ‘stops FDA from properly monitoring market-approved drugs’. Such expressions and idioms are open for interpretation, making it difficult to quantify the informativeness of a summary.

Figure 2 shows sample pairs by Subjects 1 and 4 for a non-factual (opinion) news story. Non-factual stories are not very common, but often involve a seasonal topic

2 A complete set of short and intermediate length summary-questionnaire pairs are made available at http://www.dcs.shef.ac.uk/spandh/projects/s3l/summaries.html.
such as ‘Valentine’s Day’ for this example. One noticeable aspect is the amount of abstraction preferred by the two subjects. Subject 1 fully utilised words from the news story and made a small amount of abstraction. In particular, a person name (‘Fisher’) who conducted the study was picked out, providing a backbone for the summary. On the other hand, Subject 4 has rendered the interpretation of the story in the subject’s own expressions. It resulted in a highly abstracted summary reflecting the sense of the story while ignoring the specifics. Without using the person name (‘Fisher’) it nevertheless indicated the existence of a research study behind this story by using the expression ‘science of romance’. Both summaries happened to be of good quality; however, it was the sheer divergence in the words, the expressions and subjective interpretation that was striking.

3 Cross-comprehension test

The quality of a summary can be evaluated along the GAP principles (Section 1). Some features of GAP such as reduction and organisation of facts can be supervised by defining a task (e.g., generation of a short summary in the subject’s own words). Extracting the most relevant facts from a document can be very subjective, and quantifying a complex aspect such as subjectivity in a pool of summaries is an abstruse task.

3.1 Crossing summary-questionnaire pairs

In principle, each question should be able to extract a relevant answer from a particular summary authored by the same subject. If questioned against a different summary, some answers may still stand relevant while others do not. The cross-comprehension test achieves this by swapping a summary-questionnaire pair, i.e., each summary is paired along with questions produced by different authors.

Judges (different people from the summary authors) may examine whether each question can be answered by reading a swapped summary. Further, if an answer is found, it may be relevant, partially relevant or totally irrelevant to the one expected by the question author. Thus, the decision is made from the following four options:

- **relevant**: a relevant answer is found – the answer is deemed to be relevant if it is the same in its sense as expected by the question author even if a different expression is used;
- **partially relevant**: an answer is partially relevant;
- **irrelevant**: an answer is found, but is totally different from what is expected by the question author;
- **not found**: no answer is found.

Sample (re-visited). In Figure 3, a short summary authored by Subject 2 is matched with questions set by Subject 1. By comparing the answers sought by the question author and the one found after reading the crossed summary, we may conclude answers are relevant, irrelevant, not found, not found and not found. For this particular example, the first, and probably the second, question sounds sensible to judges who
On the subjectivity of human-authored summaries

3.2 Procedure

One potential problem of this scheme is the difficulty judges may face when choosing from the four options (i.e., relevant, partially relevant, irrelevant and not found). Questions may not always be formed well for them to make a clear decision. Further, the decision process can also be a product of their own subjectivity. Our assumptions are that (1) because there are only four options, there is less room for the subjectivity in comparison to the summary writing task and that (2) a decision between relevant and partially relevant and one between irrelevant and not found are both not very important for the aim of this study because the former two are roughly associated with commonly shared information and the latter two correspond to the subjective part.

Short summaries and intermediate length summaries were individually evaluated. Each question set was paired with five different summaries produced from the same story. There were five sets of questions for each of the 51 news stories, making the total number of summary-questionnaire pairs 1,275 (5 × 5 × 51). As a consequence, evaluation was conducted for 1,275 short summary-questionnaire pairs, followed by 1,275 intermediate length pairs. Within each of the 1,275 pairs, pairs were randomly ordered, aiming at reduction of potential bias caused by memorising summaries and questions that appeared earlier. Further, the identities were masked – that is, judges were not provided with information regarding the authorship of a summary and of questions.

Ill-framed questions. Out of five summaries, four summaries are produced by the subject different from the question author. One summary, written by the question author, facilitates to measure the quality of questions. Using the summary-questionnaire pairs authored by the same subject, 95–99 per cent of answers were found relevant by three judges. The numbers improved if partially relevant cases were included. Questions marked as irrelevant or not found were considered ill-framed.

Fig. 3. Crossing summary-questionnaire pairs of Figure 1 – Subject 2’s summary is matched with Subject 1’s questions. Answers in (a1/a2) indicate the one sought by the question author (a1) and the other found after reading the crossed summary (a2). For the latter, ‘?’ implies that the answer is not found.
Table 2. Agreement between three judges for 125 short summary-questionnaire pairs

|                | \( \hat{\kappa} \) |
|----------------|---------------------|
| before discussion | 0.69                |
| after discussion  | 0.80                |

Only 1.6 per cent of questions were found ill-framed, indicating good quality of summary-questionnaire pairs produced by five subjects.

3.3 Agreement between judges

Three judges (A, B and C) conducted the evaluation. In order to improve the agreement between the three judges, the following steps were taken: initially judges evaluated 125 short summary-questionnaire pairs without any particular instruction. Later, they discussed the disagreement between them to find a common ground, which resulted in the evaluation guideline, outlined in Appendix A. It is essentially a collection of examples suggesting what decision can be made for ambiguous cases. It also includes a suggestion about what kind of common knowledge can be used, in addition to the facts found in the particular summary. Finally judges worked on the full set of 1,275 short and 1,275 intermediate length summary-questionnaire pairs. The initial 125 short summary-questionnaire pairs were repeated to measure the improvement.

The \( \kappa \) (kappa) statistic can be used to assess the level of agreement between judges (Cohen 1968). It is a measure of agreement, taking into account the agreement one would expect to see arising from pure chance. It can be estimated by

\[
\hat{\kappa} = \frac{\hat{P}(A) - \hat{P}(E)}{1 - \hat{P}(E)}
\]

where \( \hat{P}(A) \) is the estimate of the proportion of inter-judge agreements and \( \hat{P}(E) \) is the estimate of the expected proportion of chance agreements. \( \hat{\kappa} \) is ‘1’ when all subjects are in perfect agreement and ‘0’ when there is only a chance agreement.

Table 2 shows the \( \hat{\kappa} \) values between the three judges for 125 short summary-questionnaire pairs. \( \hat{\kappa} \) was 0.69 before the discussion, indicating a good agreement even without any instruction guidelines. It improved further to 0.80, or a very good agreement, after the discussion. As an additional note, \( \hat{\kappa} \)’s were 0.81 and 0.83 for the entire set of short and intermediate length summary-questionnaire pairs respectively, maintaining a very good agreement throughout the course of evaluation.

Figure 4 shows the distribution of decisions made by the three judges. The most notable change from ‘before’ to ‘after’ the discussion may be the fall of relevant, paired with the rise of irrelevant decisions. We surmise that this is the direct consequence of the established guideline that restricts the use of general knowledge at minimal. For example, ‘German chancellor’ would be the irrelevant answer when ‘Angela Merkel’ was expected, although it was certainly relevant if the greater scale of world knowledge was used.
4 Results

A good level of agreement between the three judges has been established. In this section, the subjectivity of short and intermediate length summaries is evaluated whereby average scores by the three judges are presented.

4.1 Summary relevance

Figure 5 shows, when paired with questions by other subjects, how many answers could be found in a candidate summary. The figure indicates that summaries authored by the different subjects contained relevant information for 34 per cent (short summaries) and 42 per cent (intermediate length summaries) of questions on overall average of five subjects. The number went up slightly (45%/53%) if partially relevant answers were included. These numbers provide the commonly shared part of information between subjects. The number of answers that were not found indicates the level of subjectivity for this ‘summary writing’ exercise; nearly a half (44%/41%) of information that one subject thought was the most important was discarded by the others. We observed that irrelevant answers (11%/6%) were caused either by the subjectiveness of the summary authors or by the ambiguity of the questions. Occasionally different authors arrived at summaries of inconsistent contents from each other. In such cases, questions were produced from that author’s subjective view and they certainly affected the relevance of a summary by the other subject. On the other hand, the ambiguity of a question sometimes led to a different interpretation of a candidate summary, and thus the irrelevant answer was drawn.

The numbers of relevant or partially relevant answers found in short summaries by Subjects 2 and 4 were lower (38%/40%) than others. This seems to be the consequence of a highly abstractive nature of many short summaries produced
Fig. 6. The number of answers found relevant or partially relevant was roughly proportional to the average length of summaries for each of five subjects.

Fig. 7. Questionnaire relevance – when paired with summaries by other subjects, how many questions could be answered.

by these two subjects. It is also observed that, for all subjects, the numbers of relevant or partially relevant answers found for intermediate length summaries (54%/51%/61%/55%/45% for Subjects 1, 2, 3, 4 and 5) were higher than those for short summaries (52%/40%/55%/37%/43%). This issue will be discussed further in Section 5.

Another notable outcome of this experiment is that the number of answers found relevant or partially relevant was very roughly proportional to the average length of summaries by each subject (Figure 6). The longer the summary, the more information one can write in the summary. It is thus hypothesised that the summary length certainly matters when finding the relevant information in human-authored summaries. Looking at this outcome from a different perspective, there is not much evidence that one author was more subjective than the others.

4.2 Questionnaire relevance

Figure 7 shows, when paired with summaries by other subjects, how many questions could be answered. The statistics were derived from the same evaluation as Figure 5, but observed from a different angle. Approximately the same number (39–42%) of relevant or partially relevant answers (i.e., shared information) were found for questions by Subjects 1, 2 and 3 when summaries were short. It was higher (52–54%) for Subjects 4 and 5. For intermediate length summaries, questions by Subjects 2 and 3 produced the lower number (46–49%) of relevant or partially relevant answers than those by Subject 5 (63%).
It was observed that the questionnaire relevance scores were higher for subjects who frequently set a question that might accept a wider range of answers. The lower scores were associated with subjects who tended to frame questions that required more specific information in the summary. For example, Subject 4’s ‘what is being discussed?’ was a general, open ended question that was more likely to have some answer than Subject 1’s question ‘who is the author of this study?’ Because of many open ended questions by Subject 4, the numbers of relevant and irrelevant answers have both increased for short summaries. On the other hand, Subject 5 wrote many questions focused on exact facts from an original story, resulting in more relevant answers, while maintaining a low level of irrelevant ones. A related discussion has been made in Section 2.

5 Discussion

In Section 4, by applying the cross-comprehension test on human-authored summaries from broadcast news stories, we have gauged the level of subjectivity among five authors. Three judges participated in the experiments to decide whether an answer to each question is relevant, partially relevant, irrelevant or not found.

Subjectivity and summary length. The overall numbers of relevant and partially relevant answers found by the cross-comprehension test were 45 per cent and 53 per cent, respectively, for short and intermediate length summaries. They account for the amount of information that was agreed by all the subjects as important. For roughly a half of summary contents, subjects had different opinions about whether they should be in their summaries, resulting in categories such as irrelevant or not found. Occasionally these categories were a direct consequence of ill-framed questions, but such questions were not frequent (at most a few percents in the entire set of questions – measured in Section 3). For most of the cases, they were caused by the subjectivity of a different individual.

We noted earlier that mostly the summary length matters and there is little evidence that one author was more subjective than the others. It is probably because, given a clear instruction about the summary length (i.e., approximately 15 and 50 words for this task), there is an upper bound for the amount of information that anyone can fit into the summary, while maintaining fluency. When the summary is short, one has to make a serious decision about which important information should go into a summary, and the decision often reflects one’s subjective thoughts. Our argument is that, assuming the subject’s effort, the amount of subjectivity was controlled by the summary length constraints rather than an individual’s nature.

The diversity of summaries caused by individual subjectivity may be alleviated by carefully drafting an instruction set. However, it probably results in a large list of instructions, and the drafting process certainly will not be straightforward. Further, it is not likely that we can ever completely remove the subjectivity from human work. Indeed, if subjectivity disappeared from human-authored summary by well-crafted instructions, it would be more like turning human activity into a mechanical process rather than a machine to simulate human work.
Fig. 8. Short summaries are often more distinct from each other than intermediate length summaries. The original story has an ID: ‘19980519.1830.1900.ABC.WNT.1186’ in the TDT-2 corpus (i.e., same as Figure 1).

One of the general observations arising from Section 4 is that short summaries tended to have more subjective contents than their intermediate length counterparts. Figure 8 compares the short and intermediate length summaries for the same story by Subjects 1 and 2. Despite the clear difference between short summaries by two subjects (discussed in Section 2), they produced intermediate length summaries with very similar contents, which is indicated by the choice of key words such as ‘understaffing’ vs. ‘low manpower’, ‘lack of resources’ vs ‘poor scientific infrastructure’, ‘FDA’ vs ‘Food and Drug Administration’, ‘prescription drugs’ vs. ‘prescribed drugs’, ‘side effects’ vs. ‘side effects’, ‘a million and a half hospitalisations’ vs. ‘1.5 million are hospitalised’, ‘100,000 deaths’ vs. ‘100,000 die’, ‘American Medical Association’ vs. ‘American Medical Association’.

Judge’s role. A non-trivial problem of the approach may be the amount of human effort needed for evaluation. Production of summary-questionnaire pairs may not be difficult, as it is based on the simple instruction set and even accepts ill-framed questions, but it still requires human time. Further, a judge’s work is not only highly labour-intensive, but it is the most critical for this evaluation scheme. A good level of agreement was found between three judges. It was further improved by establishing evaluation guidelines between judges and maintained throughout the course of experiments.

Level of abstraction. Abstraction seems to play a significant part both in summary relevance and in questionnaire relevance. It is often difficult to find answers to specific questions in abstract summaries while answers to less specific questions are almost always found. Short summary-questionnaire pairs by Subjects 1 and 4 in Figure 2 provide the typical example. After crossing the pairs, none of the questions prepared by Subject 1 can be answered by reading the abstract summary authored by Subject 4. On the other hand, it is probably not very difficult to find an answer to the less specific question (‘what is being discussed?’) of Subject 4 by reading the short summary by Subject 1 – although it does not imply that judge’s decision is also not difficult.\footnote{One answer derived from Subject 1’s short summary can be ‘Fisher’s study on seeking partners’. Judges may find it difficult to decide how much this is relevant to the model answer (‘science of romance’).}
Relevance to previous works addressing the similar problem. The outcome of the experiments in this paper supports the one envisaged by Lin and Hovy (2003), where an ability to evaluate a summary for facts at a sub-sentence level, rather than complete sentences or individual words, was investigated. They discussed the limitations of sentence selection for summarisation and inferred a lower inter-human agreement on the contents of a summary when summaries were shorter. Whilst they arrived at this result based on a unigram co-occurrence score between three human-authored summaries, our experiments were concerned with factual information derived from the question–answer style approach. Further the size of the summaries we have used is relatively small (15 and 50 words in comparison to their 100 words).

Similar results were reported by Radev, Jing, Styś and Tam (2004). They used 27 documents from six sources, evaluated by six judges; our experiments were based on five human-authored summaries for each of the 51 news stories from a single source, evaluated by three judges. They computed the utility score in order to measure inter-judge agreement at a sentence level, by alleviating the disagreements in the task of ranking sentences. Our work, on the other hand, aimed to quantify the subjectivity by measuring the relative informativeness (the similar facts) in a pool of human-authored summaries. Both the calculation of the relative utility and the cross-comprehension test required considerable human effort. They found a higher inter-judge agreement when the summary size was larger. Interestingly they also observed, for some documents, a drop in inter-judge agreement for summaries of compression rates 20–30 per cent, which was recovered for 10 per cent compression.

The factoid analysis (Teufel and van Halteren 2004) and the pyramid method (Nenkova, Passonneau and McKeown 2007) have tried to quantify the relevance of atomic facts. Both approaches involved some annotation and weighting of the atomic facts for importance. The factoid analysis investigated 50 different summaries from two news stories. Atomic facts were extracted from these summaries and defined empirically on the data by annotators. Increasing inter-annotator agreement was observed when the annotator had seen a list by the other annotators. It was argued that several summaries were needed to create a full factoid inventory. Once the factoids were listed, they were rated for significance based on their frequency and location. Factoids outperformed the unigram in the task of summary evaluation. In comparison, the method discussed in this paper does not rate summaries created; it is the number of questions that cannot be answered that reflects the subjectivity of a summary.

Using the pyramid method, Nenkova et al. (2007) analysed the human abstracts at a sub-sentential level to quantify the human variation. They used summary content unit (SCU), which was roughly an atomic fact, as the unit of comparison. An $n$-tiered pyramid was formed where $n$ was the number of model summaries. The SCUs were then weighted according to the occurrence in the number of model summaries. It was claimed that such weighting was necessary in summary evaluation because each individual might choose somewhat different information when requested to write a summary for the same set of documents. The pyramid scores relied on creation of the pyramid and annotation of peer summary against pyramids. They also found an increased inter-annotator agreement when the individual annotators had the benefit of annotations made by peers. In comparison, the work in this paper tries to quantify
the subjectivity in various summaries in terms of interpretable answers. This is done by examining a candidate summary to see if it contains the information sought by the referred question. Using an example, further comparison is made in Appendix B to illustrate the similarity and some differences between the pyramid method and the cross-comprehension test.

In summary, while most other approaches appear to assign a significance score to the atomic facts, the scores resulting out of the cross-comprehension test rely on the ability to deduce an answer for a given question from a candidate summary. The cross-comprehension approach pans out on a different dimension than ranking the atomic facts. The answers for the respective questions are given by the summary authors and represent their perspective. Hence they are beyond the scope of ranking. To limit the subjectivity, our approach relies on the judges’ ability to interpret a plausible answer with reference to the viewpoint of the summary author.

Cross-comprehension test may seem analogous to question-answering evaluation; however, the principles underlying the two are rather different. While the former identifies the information commonly shared amongst different summaries, the latter tries to find an optimum answer from a variety of answers available. Further, the former evaluates a candidate summary sought by the question author and the latter involves comparison of two or more answers to a question.

**Potential for relative evaluation.** The ability for measuring the subjectivity implies that the approach is able to discard the subjectivity when evaluating the informativeness of candidate summaries. The major benefit of crossing summary-questionnaire pairs is that the approach avoids absolute values, but it relies on a relative comparison. Ill-framed questions and the ambiguity caused by crossing summary-questionnaire pairs could be cancelled out using relative comparison.

Figure 9 shows the outcome of relative evaluation, where short summaries by Subjects 1, 2 and 3 are matched with questions authored by Subjects 4 and 5. While most of the questions by Subject 4 focused on an exact content of a summary, Subject 5 wrote many open-ended questions (noted in Section 2). The figure shows that they resulted in different amount of answers that were relevant, partially relevant, irrelevant or not found. Despite the above difference, it was found that summaries by Subject 3 (or 2) share the most (or least) amount of common information with Subjects 4 or 5.

The relative evaluation was initially tested in Kolluru, Christensen and Gotoh (2005) and further investigated in Kolluru and Gotoh (2007). The approach reflects relative informativeness of the candidate summaries without treating any specific
summary as a ground truth. The subjectivity can be a teething issue for summary evaluation. However, there is a way around it by crafting an approach, such as the cross-comprehension test, that makes a relative comparison.

6 Conclusion

In recent years, we have seen a steady progress in automatic production of summaries in various settings, whereby summary evaluation has been recognised as a sensitive, non-trivial task. For quite sometime now a produced summary is measured against a reference, which is typically authored by a human and referred to as a ‘gold standard’ (Mani 2001). One unavoidable, but largely overlooked, aspect of human-authored summaries is their subjectivity. Consequently, no matter what approach is adopted for evaluation, there always exists a concern if they really reflect the quality of candidate summaries.

This paper presented one approach to measure human subjectivity quantitatively when authoring summaries of short to intermediate length. Applying the cross-comprehension test on human-authored summaries for broadcast news stories, the level of subjectivity is analysed among five subjects. We are aware that the experiments might not have shown the complete picture. Even with the same instruction set, short summaries tended to be more subjective than their intermediate length counterparts; then what if the different instructions would be given for summary authors? What if the approach would be applied for the different tasks, or for the different domains? Although the issue of subjectivity may require further investigation, the approach presented can be applicable under various scenarios.

References

Boret, M. and Peyrot, J. 1969. Le Résumé de Texte: Épreuve d’une Pédagogie Nouvelle. Paris: Chotard et Associés.
Cieri, C., Graff, D., Liberman, M., Martey, N. and Strassel, S. 1999. The TDT-2 text and speech corpus. In Proceedings of DARPA Broadcast News Workshop, Herndon, VA, pp. 57–60.
Cohen, J. 1968. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 20: 37–46.
Cremins, E. T. 1982. The Art of Abstracting. Philadelphia, PA: ISI Press.
Dang, H. and Lin, J. 2007. Different structures for evaluating answers to complex questions: pyramids won’t topple, and neither will human assessors. In Proceedings of the ACL 2007, Prague, pp. 768–775.
Endres-Niggemeyer, B. and Neugebauer, E. 1998. Professional summarizing: no cognitive simulation without observation. Journal of the American Society for Information Science 49(6): 486–506.
Hirschmann, L., Burger, J., Palmer, D. and Robinson, P. 1999. Evaluating content extraction from audio source. In Proceedings of the ESCA Workshop: Accessing Information in Spoken Audio, Cambridge, UK, pp. 54–59.
Hovy, E., Lin, C.-Y., Zhou, L. and Fukumoto, J. 2006. Automatic summarization evaluation with basic elements. In Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006), Genoa.
Jing, H. 2002. Using hidden Markov modeling to decompose human-written summaries. Computational Linguistics 28(4): 527–543.
Kintsch, W. and van Dijk, T. A. 1978. Toward a model of text comprehension and production. *Psychological Review* **85**(5): 363–394.

Kolluru, B., Christensen, H. and Gotoh, Y. 2005. Multi-stage compaction approach to broadcast news summarisation. In *Proceedings of Interspeech 2005*, Lisbon, pp. 69–72.

Kolluru, B. and Gotoh, Y. 2007. Relative evaluation of informativeness in machine generated summaries. In *Proceedings of Interspeech 2007*, Antwerp, pp. 1338–1341.

Lin, C.-Y. and Hovy, E. 2003. The potential and limitations of automatic sentence extraction for summarization. In *Proceedings of the HLT-NAACL 2003 Workshop on Automatic Summarization*, Edmonton, pp. 73–80.

Lin, J. and Demner-Fushman, D. 2006. Will Pyramids built of nuggets topple over? In *Proceedings of HLT/NAACL 2006*, New York, pp. 383–90.

Mani, I. 2001. *Automatic Summarization*. Jon Benjamins Publishing Company, Amsterdam.

Nenkova, A. and Passonneau, R. 2004. Evaluating content selection in summarization: the pyramid method. In *Proceedings of the HLT-NAACL 2004*, Boston, pp. 145–152.

Nenkova, A., Passonneau, R. and McKeown, K. 2007. The Pyramid method: incorporating human content selection variation in summarization evaluation. *ACM Transactions on Speech and Language Processing* **4**(2), article 4.

Pinto Molina, M. 1995. Documentary abstracting: toward a methodological model. *Journal of the American Society for Information Science* **46**(3): 225–234.

Radev, D. R., Jing, H., Styś, M. and Tam, D. 2004. Centroid-based summarization of multiple documents. *Information Processing and Management* **40**, 919–938.

Rowley, J. E. 1988. *Abstracting and Indexing*, 2nd ed. London: Bingley.

Teufel, S. and van Halteren, H. 2004. Evaluating information content by factoid analysis: human annotation and stability. In *Proceedings of the EMNLP 2004*, Barcelona, pp. 419–426.

Van Halteren, H. and Teufel, S. 2003. Examining the consensus between human summaries: initial experiments with factoid analysis. In *Proceedings of the HLT-NAACL 2003 Workshop on Automatic Summarization*, Edmonton, pp. 57–64.

Vanderwende, L., Banko, M. and Menezes, A. 2004. Event-centric summary generation. In *Proceedings of DUC 2004*, Boston, pp. 76–81.
Appendix A: Evaluation guideline for judges

1. Try to answer the question rather than to find its answer in the summary.

notes: This means that a judge should answer the question for the summary without looking at the answer given by original question author. If an answer is found then the judge may compare it with the original and decide whether it is relevant, partially relevant or irrelevant. The rule would elicit information that is expected precisely by the summary author rather than a pattern matching exercise of searching answers.

examples:

‘Who is the president of America?’ (Bill Clinton)
If the summary has ‘US president Clinton’, then the answer would be relevant. If the summary is like ‘Clinton denied having …’ then the answer would be not found, because the question cannot be answered based on the information provided in the summary.

‘Where did the earthquake occur?’ (Afghanistan)
If the summary has ‘Tuesday’s earthquake in Afghanistan …’ then the answer is relevant, but if the summary says ‘… terrible earthquake. Afghanistan is unable to cope …’ then the answer is likely to be not found as the question cannot be answered from the summary.

2. Assume the previous question(s) have been accurately answered when disambiguating a question.

notes: This rule would enable a judge to treat each question independently. Without this rule, the evaluation would be skewed towards initial questions.

examples:

‘Who was at the negotiating table?’ (Iraq and UN)
‘Were they successful?’
The second question can be read as ‘were Iraq and UN successful at the negotiation?’

3. Use only a basic level of general knowledge.

notes: Given the vast un-imitable human knowledge, we had to restrict the knowledge sources to the news summary concerned.

examples:

‘US$’ or ‘dollar’ is the currency of the United States.
‘A few thousand miles’ is a ‘considerable distance’.
‘German Chancellor’ may not be ‘Angela Merkel’.

Appendix B: Pyramid method and cross-comprehension test

The purpose here is to illustrate the similarity and some differences between the pyramid method and the cross-comprehension test, but it is not our intention to achieve a full-scale comparison. We implemented the pyramid based on our interpretation of Nenkova et al. (2007). A single annotation was made for the intermediate length summaries derived from a story.
Fig. B1. We tested the pyramid method using five intermediate length summaries derived from the same story as Figures 1 and 8. We found one, three, one and two SCUs with weight 5, 4, 2 and 1, respectively. Thus the total weight of the ideal summary would be $T_{\text{max}} = 21$.

**Subject 1:** [pyramid score: 0.81]
Due to understaffing and lack of resources (‘lack of resources’), the FDA is failing to thoroughly vet prescription drugs for public consumption (‘FDA failed’). Side effects from some of these drugs cause a million and a half hospitalisations and 100000 deaths yearly (‘side effects, some deaths’). American Medical Association recommends creation of independent drug safety office to assume FDA drug vetting responsibilities (‘action necessary, new commissioner’).

**Subject 2:** [pyramid score: 0.81]
American Medical Association in a study on Food and Drug Administration noticed inadequate performance (‘FDA failed’) as 1.5 million are hospitalised and 100000 die due to side effects of prescribed drugs (‘side effects, some deaths’). FDA is struggling with poor scientific infrastructure, low manpower and poor drug monitoring systems (‘lack of resources’). Clinton might appoint an effective commissioner from Mexico (‘action necessary, new commissioner’).

**Subject 3:** [pyramid score: 0.76]
1.5 million Americans hospitalised by prescription drug side effects each year. 100000 die (‘side effects, some deaths’). American Medical Association says Food and Drug Administration which is required to monitor drugs not doing its job (‘FDA failed’). Backlog of 100000 drug reports (‘backlog’). 52 people follow almost 200000 prescription drugs (‘52 using 200000 drugs’). Clinton may appoint more effective FDA commissioner (‘action necessary, new commissioner’).

**Subject 4:** [pyramid score: 0.81]
FDA accused of not properly monitoring prescription drugs for harmful side effects after approving them (‘FDA failed’). FDA lacks funding and mandate to carry out monitoring (‘lack of resources’). Each year one point five million people hospitalised because of side effects from prescription drugs; one hundred thousand die (‘side effects, some deaths’). Commentators call for a new program to monitor drug safety (‘action necessary, new commissioner’).

**Subject 5:** [pyramid score: 0.57]
The Food and Drug Agency, responsible for ensuring the safety of medication in America is not doing its job properly (‘FDA failed’). There is a backlog of reports of bad medication (‘backlog’) partly because of a poor computer system (‘lack of resources’). The FDA is under pressure to get drugs to the marketplace quickly (‘FDA under pressure’).

Fig. B2. Using the list of SCUs in Figure B1, pyramid scores were calculated for five intermediate length summaries. Items in ( ) indicate the corresponding SCU to the preceding phrase. For example, four SCUs with weight 4, 5, 4 and 4 were found in the summary by Subject 1. Thus the accumulated weight is 17 and the pyramid score is $17/21 = 0.81$.

Firstly, a set of SCUs was annotated at a phrase level based on their occurrence in the summaries by five subjects. SCUs were weighted against their respective occurrences in the summaries (Figure B1). The ideal summary score $T_{\text{max}}$ was the sum of all distinct SCU weights. Secondly, we calculated the pyramid score for individual summaries as the ratio $D/T_{\text{max}}$, with $D$ being the entire weights recorded by the presence of SCUs in that summary (Figure B2). In this particular example, the higher pyramid scores for the summaries created by Subjects 1, 2, 3 and 4 indicate that they are closer to the ideal summary than the one by Subject 5.

For the same set of intermediate length summaries, Figure B3 compares the scores by the pyramid and the summary relevance by the cross-comprehension test. They have roughly a linear relationship; although the scale of comparison is too small to draw a definite conclusion, the cross-comprehension test appears to support the measure used in the pyramid. According to our interpretation used in this paper, a summary by Subject 5 has less agreement with the others, i.e., it is more subjective than the rest. The pyramid revolves around the formation of an ideal summary, or ‘gold standard’, while the cross-comprehension test is about information pool which
identifies similarities and differences among candidate summaries. It is thus not our intention to rate individual summaries in a way to decide which summary is better. In the pyramid method, annotation of SCUs is made by a third person, while in the cross-comprehension test questions are formed by a summary author. This leads to another conceptual difference between two schemes; in the former, annotators make an exhaustive search for SCUs presented in a summary and occurrences of individual units are treated equally. In the latter, the summary author has a chance to pick the most important facts that he/she thinks should go into summaries. In this paper, we have presented two types of relevance scores; the summary author’s intention may be more evident in the question relevance than in the summary relevance. Finally, facts in the cross-comprehension test look more granular than SCUs in the pyramid although this may not be of fundamental importance.

![Graph](image-url)

Fig. B3. For five summaries in Figure B2, the number of answers found relevant or partially relevant was roughly proportional to the pyramid score.