

# Evaluating expected and actual success of automatically generated summaries and entity extraction

Anonymous Authors  
Department  
Institution  
Email@address.suf

## ABSTRACT

After ten years of research, automatic summarisation techniques are now maturing and the arrival of APIs and freely available plugins means that they can be easily integrated into any information system. Even though there has been extensive evaluation, using benchmarks and gold-standard comparisons, research has yet to focus on utility, usability, and human opinion of whether a summary is acceptable and successful or not. Using a two-phase repeated measures user study, we examined the expectations and reactions of everyday computer users to human- and automatically-generated summaries. Rather than benchmarking and comparing possible techniques, we take the first steps to understanding what everyday computer users think about automatic summaries, and in comparison to human summaries. Our results show [TODO – or While participants in phase 1 believed x and y...]. We conclude that [A and B], which has design implications for [blah and blah]. While our results show [major discover], further work is required to evaluate automatically generated summaries, and perhaps entity sets, support users in scenario-driven task-based user studies.

## Author Keywords

Automatic Summarisation, Entity Extraction, Expectations.

## ACM Classification Keywords

[TODO]H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

## General Terms

[TODO]See list of the limited ACM 16 terms in the instructions, see <http://www.sheridanprinting.com/sigchi/generalterms.htm>.

## INTRODUCTION

The methods for automatically summarising one or more pieces of text are maturing, boosted by algorithmic developments made in natural language processing. Over the past ten years, there have been a number of workshops and symposiums dedicated to improving the tools that

summarise text automatically. During this time, evaluating the success of summarisation algorithms has been a technical process giving praise to those algorithms which score well in terms of presence of key words and entities, rather than those that have actual utility and the coherence of a good piece of text. Using gold-standard summaries for a given text as a standard, evaluation has typically involved benchmarking and comparing approaches. This evaluation technique is similar to the Cranfield and TREC evaluations used by the Information Retrieval community.

As automatic summarisation tools and the underlying technologies have improved, openly available APIs and downloadable plugins mean that automatic summarisation can now be easily embedded in any software. Microsoft, for example, included a free summarisation tool in Office 2007<sup>1</sup>. Consequently, we now need to know more about the actual utility of summarisation algorithms, whether they are easily accepted by everyday users, and how they are then used. Understanding these factors may explain why summarisation was removed from Office 2010 [1, 2], and whether it has more appropriate utility in other applications.

This research has taken the first steps towards addressing these unknowns about the utility and acceptance of automatic summarisation. The first steps presented here have focused on understanding user expectations about automatic summarisation, and reactions to how successful they are at representing larger documents. After describing more about automatic summarisation and related work, we describe a two-phase study that captures expectations in Phase 1 and reactions and success-judgements in Phase 2. We conclude with a discussion of implications and design recommendations, and describe the necessity for future work in this area that evaluates summarisation technologies in both lab-based and real-world studies.

## RELATED WORK

Automatic summarisation aims to help people quickly make sense of large portions of text, or to identify likely sources that contain useful information. Consequently, the related work below begins by framing automatic summarisation in

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

CHI 2012, May 5-10, 2012, Austin, TX, USA.

Copyright 2012 ACM xxx-x-xxxx-xxxx-x/xx/xx...\$10.00.

---

<sup>1</sup> <http://office.microsoft.com/en-us/word-help/automatically-summarize-a-document-HA010255206.aspx>

the context of searching; for resolving a gap in knowledge or an information need.

### **Information Retrieval/Seeking in Search**

Information retrieval (IR) is a large and mature field of research. IR tries to satisfy a user's information need, expressed as a query in natural language form, from a potentially limitless collection of documents [1, 3]. Search engines have developed increasingly efficient and ingenious ways to find results that match, or are relevant to, a submitted query. IR, however, represents only a small portion of the overall information seeking process, where Information Seeking (IS) is defined as resolving an information need [4]. Consequently, the IS process involves a user: recognising that they have an information need, understanding their need, performing one or more keyword searches (IR), assessing results, reflecting on information, and eventually finishing once the information need has been resolved. Where search engines have been well researched for returning results for queries, approaches like automatic summarisation (discussed below) and information visualisation [3] [5] aim to help users to better assess results and reflect on information.

Whether a search engine is successful is complicated due to the complex mixture of both subjective and statistical measures. In the early days of web search, Witten outlined the success of searching as how pertinent items can be located and information extracted without undue expense or inconvenience [6]. Therefore, in order for a search to be successful, not only is IR about locating the most relevant documents, but it is also about doing so in the quickest possible manner. In 1998, the Google's average search-response was anywhere between one and ten seconds [7], however better indexing techniques have further improved this to the millisecond responses we have today [8].

The efficiency and accuracy of IR algorithms has thrived over the last few decades because of the Cranfield evaluation paradigm [9] and subsequent TREC conferences<sup>2</sup>. These evaluation frameworks use known gold-standard relevant documents, from a given corpus and for a set of given queries, so that algorithms can be benchmarked and compared quickly and easily. As these techniques have developed, and search engines have implemented such improvements, accuracy and efficiency are now rarely unique selling points for search engines.

The success of different search services has gone beyond IR, especially within vertical search environments like online retail, to supporting the larger IS process. Microsoft studied the factors that have made search engines successful [10]. They performed a study to compare the re-usage of both major and minor search engines between stationary and non-stationary uses to understand the likeliness of people returning to an environment over another. Their

results confirmed that for a search engine to be successful they not only need to concentrate on convincing users to try their search environment but also convincing them to continue using their search engine over competitors [10]. They also established that those search engines who don't satisfy their users over information needs will more than likely see a reduced market share [10]. Therefore it is important that search engines put the solving of the information need as the focus of their system.

Companies specialising in web searches have modified their approaches in many ways to support people throughout the IS process [11]. Much research has focused on providing useful search suggestions, with Ruthven et al [12] showing that search engines who implemented search suggestions lead to better search results in comparison with the control. Google now tries to help users to produce better queries by showing possible queries *and* the first results while the user is still typing their query [8]. Search engines also help people to identify more useful results by highlighting results that they have seen before, or that friends have recommended [13]. Collaborative filtering techniques for recommender systems (e.g. [14]) use the behaviour of similar users in order to recommend results that searchers might be interested in. There are many countless other techniques, such as filtering, sorting, and clustering, that have been developed to support interactive IR, and the IS process as a whole. Not surprisingly, much of the evaluations of systems that support the broader IS process are focused on the utility and usability of search user interface techniques, rather than on improvements on speed and accuracy alone.

### **Automatic Summarisation Principles**

Automatic summarisation is defined as taking an information source, extracting content, and presenting the most important content to the user in a condensed form in a manner sensitive to the user's or application needs [15]. Consequently, automatic summarisation helps to support the latter phases of the IS process, by helping searchers to analyse results and reflect quickly on the value of an information resource. Automatic summarisation, therefore can help to satisfy Witten's success measure [6], as not only does it present the required information to the user but does so in a way which is easy for the searcher to access. While several techniques for automatic summarisation have been developed, however, the majority of research has focused on the accuracy and efficiency of algorithms, in terms of how well they pick out successful/relevant terms, phrases and paragraphs [16], rather than their utility within search and how successful users believe they are.

Summaries developed by computer algorithms typically conform to two main techniques. The majority of algorithms use the topic identification [17] (aka extraction [18]) method, which works by modifying well-used ranking techniques such as TF-IDF to rank words, sentences, and paragraphs by their relevance [17]. This approach uses

---

<sup>2</sup> <http://trec.nist.gov/>

statistics to give us a best estimate on which sections are more relevant, so that the computer can compile a summary out of the top ranked sections [17]. This technique is most common due to its ease of implementing a base line algorithm. Summaries based on extraction, however, can often lose important contextual detail, and their flow can be poor due to language complexities.

Less commonly, algorithms can also be based on interpretation (aka topic fusion [17] or abstraction [18]) that builds on the topic identification extraction phase. Interpretation works by attempting to formulate sentences using words and phrases extracted in the topic identification phase reworked into a new text formulation using new words and phrases not present in the original source [17]. This technique is less common as it requires algorithms to be able to process natural language and understand the scope of the text as a whole, developing the summary automatically, which requires a high degree of artificial intelligence [17]. Even though topic interpretation is rare, the algorithms being used in natural language processing are becoming more and more complex [2], which supports the prediction that abstraction will become more common within automatic summarisation tools as time progresses.

#### **Evaluation of Automatic Summarisation Tools**

Until now, most of the research within the field of automatic text summarisation has involved firstly understanding what makes a good and a bad summary, and using this information to evaluate algorithms. A common question that has dominated all such comparative evaluations within this field, however, is how to rate the success of one summary over another. What the ‘best’ summary should include is an entirely subjective matter [18] and will not only change from person to person but is also likely to change from time to time for the same person [18]. This is because the opinions of what a user might feel is relevant can vastly differ due to limitless reasoning including differing expertise, language competency, profession and semantic interpretation [19]. Not only does this mean that people interpret text differently, it also means that evaluation of algorithms and systems is difficult as evaluation usually consists of humans opinions. Both of these reasons combined mean that there is no 100% successful way of summarising text [19].

To combat the effect of differing human opinions, evaluations tend to use benchmark comparison with a ‘gold-standard’ summary written by human experts or computer algorithms [20]. Notably, this approach is much like the early Cranfield and on-going TREC conferences used to develop better IR algorithms. Researchers within this field, however, have not only had to address how to evaluate a summary but also *who* evaluates the summary. The issue on humans being subjective once again arises, as assessors will have differing opinions on the success. Human evaluation was the chosen method by the Document

Understanding Conference (DUC)<sup>3</sup> from 2001 – 2003 but was changed to computer evaluation from 2004 onwards [21]. Computer evaluation also has problems, however, as the success of the evaluation is only as good as the algorithms that are designed and it is likely to fail to take into account context and coherence.

The turn of the century saw a surge in activity relating to automatic summarisation with a number of AS dedicated workshops such as the DUC<sub>3</sub> and a dedicated TREC conference<sub>2</sub> boosted by the advances in the natural language processing (NLP) field. Despite this activity, progress was limited mainly due to the lack of standardisation of evaluation meaning that it was difficult to assess the progress due to different people using vastly different success measures. Lin and Hovy attempted to resolve this by evaluating the evaluation techniques which were common at the time [16]. They found that human evaluation traditionally performed well against automatic scoring techniques, however they felt that more resources should be dedicated to the production of automated evaluation systems [16]. Subsequent work led to the development of “ROUGE”, a system that was used to automatically evaluate a summary against a gold-standard using statistical measures such as n-gram, word sequences and word-pairs which were common in both which was adopted as the official evaluation package for the DUC [21].

Even though ROUGE has been adopted as the evaluation standard, it is far from a perfect evaluation technique. Even though it spots common words, phrases and sentences across the two summaries, it has no support for synonyms or paraphrases of the words and/or phrases included in the passage [21]. Therefore, one sentence might have the same meaning but be given a lower rating, because they are lexically different. ParaEval found that by using translation tools and spotting those words or phrases that had the same Chinese output from the English input, they could interpret what were paraphrases of one another [16]. By adding this technique to the top layer of the ROUGE framework, they could guarantee improvements to the success rating culminating in a 0.035 Pearson improvement against the ROUGE on the DUC 2003 conference.

#### **Automatic Summarisation in Practice**

Even though there has been extensive work on the creation and evaluation of algorithms that specialise in summarising a piece of text automatically, comparatively there has been relatively little work into the actual implementation of automatic summarisation into situations that a user can use. As previously discussed, automatic summarisation aims to extract key information and present it in a way that suits the user’s needs. The success of the summary based on the ROUGE framework, however, has no scope for taking into

---

<sup>3</sup> DUC: <http://duc.nist.gov/>

account what the human needs that information for and the scenario the human is working in.

One study that attempted to take into account the scenario that the human was working in, was focused on the development of a system that created summaries of health and safety information in end of year company reports [22]. Due to the specific nature of the information need by the user, the system correctly retrieved all of the right information from the source text on 70% of occasions[22]. By producing specialised systems within a specific scenario, the issues relating to the context of the search were all but eliminated, which dramatically improved the success rating.

Even though the problem of lack of context and coherence in automatic summarisation tools still exist, and some have started making their way into the mainstream. Microsoft's AutoSummarise (although removed in Office 2010 [1]) was provided to automatically summarise documents and create executive summaries. It's not clear why this tool has since been removed, but unlike most envisioned scenarios of use, AutoSummarise was provided to help build a document. Automatic summarisation was also used as one of the unique selling points of the Cuil (re-launched as Cpedia) search engine [23], which has since been shut down [24]. Cuil/Cpedia aimed to generate Wikipedia style summaries over groups of results but was criticised widely across the industry [23]. Pingar provides a service that creates a summary PDF of the top search results, with links to the original documents. There are several tools available to be reused. Pingar provides an API<sup>4</sup> and there are many libraries that can be downloaded for use<sup>5</sup>.

The fact that these services are emerging, and have sometimes failed, further indicates that we know much less about the utility and benefits of automatically created summaries than we do about how to generate them. This study described below represents our first steps towards examining what people think about summaries, and how they compare to human summaries when trying to understand unfamiliar documents.

#### **USER STUDY – DETERMINING THE SUBJECTIVE BASED NATURE OF SUMMARIES**

The aim of automatic summaries, in helping users to examine results or documents by synthesising and summarising them, is clear, but we have little understanding about their actual utility and how users respond to them. We were motivated by understanding these issues better and consequently our research questions were:

RQ1: What are peoples' expectations about computer-generated summaries?

RQ2: Can people tell whether summaries are computer-generated?

RQ3: How successful are computer-generated summaries at helping users understand a topic?

Unlike most prior research, our aim was not to look at performance improvement, but how they affect and support the IS Process. In the long run, however, such analyses would involve measuring success according to changes or differences in the complex nature of learning and sensemaking. Understanding how summaries affect performance or behaviour are valuable paths for future work, but as a first step, the work presented here is initially concerned with understanding how human users perceive summaries, and with analysing judgements of their quality. Consequently, we generated the following hypotheses.

H1: Participants will expect computer-generated summaries to be significantly different from human-generated summaries.<sup>6</sup>

H2: People will be able to accurately tell whether a summary is computer-generated or not.

H3: Human- and computer-generated summaries will differ significantly in quality.

H4: Human-generated summaries will be significantly more useful as they increase in size.

H5: Computer-generated summaries will be significantly more useful as they increase in size.

To test these hypotheses, we performed a study in two phases that covered both predictive expectation and then actual empirical judgements of summaries. Phase 1 was focused on H1, but had two aims: a) gathering human-generated summaries (to be used in phase 2), and b) evaluating participant expectations around automatically generated summaries. Phase 2 was focused on H2-H5 and so had three aims: a) gathering ratings for both human- and computer-generated summaries, b) determining whether participants could accurately tell human- and computer-generated summaries apart, and c) evaluating human- and computer-generated summaries of different sizes. In both phases, the research ethics board approved the procedure, and participants received a £10 Amazon voucher as a gesture of goodwill for giving up their time.

#### **Texts and Technology**

Five texts were used in the study, which were taken from Wikipedia and the Guardian newspaper and had a mean length of 755 words. The five texts were on: the game of Kubb (T1), the Watership Down novel (T2), the Bloodhound vehicle (T3), the Edinburgh Fringe festival

---

<sup>4</sup> <http://www.pingar.com/DevelopersRD.aspx>

<sup>5</sup> e.g. <http://libots.sourceforge.net/>

---

<sup>6</sup> We left this hypothesis as two-tailed, as the two authors themselves could not agree about which type participants would expect to be better.

(T4), and history of animated movies (T5). The texts were chosen to be approximately comparable in size, technical content, and difficulty (when summarised by human participants). These texts, or their summaries, were used in both phases, but different participants took part in each phase in order to remove learning effect when empirical ratings of human- and computer-generated summaries were compared.

*“Speed enthusiasts hope to build a rocket car that can go faster than a bullet from a handgun -- and break the world land speed record. The project involves wing commander Andy Green, (a Royal Air Force fighter pilot) and Richard Noble, who also project-led the Thrust SSC and himself held the world speed record between 1994 and 1997. The project is privately financed through sponsorship, hopes to promote public interest in science and technology as well as develop new technology in engineering. The Bloodhound team are now searching the world's deserts, hunting for a race-track capable of taking the supersonic car”*

**Figure 1 - Example of summary developed by Pingar API**

As the research was not motivated by comparing and evaluating different algorithmic approaches, we sought to use a state of the art service that was representative of the openly available summarisation algorithms online. We chose the enterprise level API<sup>7</sup> provided by Pingar to generate summaries for our texts. Pingar’s API is a recent release that moves their services from integrated business-to-business packages to being open and broadly available for any conceivable application. Consequently, Pingar’s technology both confirms the notion that automatic summarisation technologies are becoming increasingly available online, as well as being representative of the technology that could be used to power the experiences that everyday users may have with websites. Pingar’s API was used as a baseline for comparison in our research, although not in Phase 1, which was focused on gathering expectations about computer-generated summaries with an example of a summary developed by the Pingar API shown in [FIGURE].

Pingar do not publish their exact methods for summarising text. The summaries produced, however, use statements taken directly from the text, indicating that they use the section-ranking approach to identify the key text to include in summaries. Further, Pingar’s core staff includes those that have worked on entity extraction in Wikipedia [25].

In both phases, a simple .NET website was generated in C# to: gather pre- and post-study survey data, control the randomisation of independent variables, control the flow of questions, to gather answers, and provide some standardised instructions and help. Both phases were performed in a quiet room, using a 13-inch dual-core

MacBook Pro and an external mouse and keyboard. Participants, however, used Internet Explorer over a remote desktop connection to the built system. Responsiveness had been excluded as a concern prior to beginning the study.

## **PHASE 1 – EXPECTATIONS AND DATA COLLECTION**

The motivation of phase 1 was two-fold. First, Phase 1 collected human-generated summaries that could be used in Phase 2. These human-generated summaries were later used to compare human- and computer-generated summaries. Second, Phase 1 collected a base-case data set of human-expectations, prior to seeing or analysing any automatically generated summaries, and based upon their own experiences of summarising text. These expectations are compared in Phase 2, against the opinions of people who have directly experienced the technology first hand.

### **Methodology**

10 participants, 7 female and 3 male, were recruited across a British university, using bulk email to all staff and students. Participants ranged between 25 and 60 years old, covering a range of professions and disciplines including administration, technicians, librarians, and research staff. The aim was to get mixed backgrounds, rather than computer scientists who may be familiar with text analysis. All 10 described themselves as frequent users of Google.

After providing informed consent, participants were asked to generate five summaries, one for each text, of different sizes. No constraining time limit was specified. Summary-size and text-analysis were the primary independent variables, and were counterbalanced to avoid the following potentially confounding variables: a) growing familiarity with the text, and b) practice in writing summaries in increasing or decreasing order of size. The five sizes used were measured in sentences, using one sentence for the smallest and five for the largest. For the task, participants were asked to write a summary of each text that would answer the question ‘What is [the topic]’, in a specific number of sentences.

After approximately 30-40 minutes of writing summaries, participants were provided with a short survey. On a 9-point Likert scale, participants were asked to rate overall difficulty, and the difficulty of summarizing each topic. Further, participants were asked to rate, again using a 9-point scale, the quality of their own summaries, and the quality that they expected computer-generated summaries would be of the same texts. Beyond collecting summaries, the system also measured the time taken to create the summaries. The session ended with a short debriefing interview, which collected additional qualitative data about their Likert-scale answers and the potential utility of automatic summaries.

### **Results**

#### *Human-generated summaries*

The main aim of this section is to describe the accumulation of human-generated summaries, and what they were like.

<sup>7</sup> <http://www.pingar.com/DevelopersCA.aspx>

Our results indicated that there was no significant difference between the texts, the tasks, and the summaries generated by human participants.

Difficult per Text Analysed	T1	T2	T3	T4	T5
Mean	4.8	4.7	5.2	4.6	4.5
STD	2.6	2.2	2.1	1.7	1.7

**Table 1: Table showing the average difficulty scores assigned to each topic.**

Overall, participants rated the average difficulty of creating the summaries as 4.3 (STD: 2.3). [TABLE] shows the average difficulty ratings for each text. A Friedman test revealed no significant difference between the difficulty of summarising each text. To further establish that all the summaries collected for Phase 2 were fair, we analysed the number of words for each size of summarisation across the texts. Again, no statistical difference was found in the sizes of summaries generated, which are shown in [TABLE].

As we found no significant differences, we concluded that both the texts used in Phase 1, and the summaries generated for Phase 2 were indeed comparable across the five topics.

Size	T1	T2	T3	T4	T5	Mean
1S	40	46	37	45	32	<b>40.0</b>
2S	40	65	28	37	43	<b>42.6</b>
3S	91	94	53	66	66	<b>74.0</b>
4S	70	71	103	69	69	<b>76.4</b>
5S	116	97	126	118	118	<b>115.0</b>
<b>Mean words per sentence</b>	<b>71.4</b>	<b>74.6</b>	<b>69.4</b>	<b>67.0</b>	<b>65.6</b>	<b>69.6</b>

**Table 2: The number of words for each size and topic of human-generated summaries**

#### *Expectations of Computer-generated Summaries*

Shortly after writing five of their own summaries, we asked participants to rate both the average quality of their summaries (avg: 5.5, STD: 2.1), and the quality of summary that they expected a computer could generate (avg: 5.3, STD: 1.6). A Mann-Whitney U test revealed that there was no significant difference between these two sets of data. Consequently, we rejected H1, as our participants expected that computers would be able to automatically generate summaries of roughly equal quality to their own.

These ratings were discussed in the post-task debriefing interview. Participants reported sticking to the middle region of the Likert scale because they felt that they couldn't have strong opinions either way. This was perhaps because they were yet to see an example of an automatically generated summary. This contrasted with summaries that participants made themselves, as the spread was much greater as the rating closely conformed to the

confidence they had in their abilities. [Why did some report higher and some lower?].

#### *Situations where Automatic Summarisation Is Useful*

We also asked participants to outline what situations common to their life would an automatic summarisation tool be a useful tool to have. As all ten participants in this phase were staff members within the university, it enabled us to see where they could use the technology in both their home life and work.

The main themes that ran throughout the situations presented were time and length. They felt that an automatic summarisation tool was perfect for those occasions where they had a wealth of information and wanted to quickly know whether to read on. For example, some of the situations mentioned included research papers/academic journals (5 mentions), business reports (2 mentions), web page summaries in search engines (2 mentions) and news articles (1 mention). No participants within this phase failed to give an example situation, which proved that people believe there are situations that an automatic summarisation tool would be useful to have. We can conclude there is a requirement for this technology and practical implementations of a text summarisation tool into web-based systems would be welcome.

#### **PHASE 2 – REACTIONS TO SUMMARIES**

While Phase 1 focused on gathering human-generated summaries and understanding expectations about automatic summarisation, Phase 2 aimed to establish some strong findings about the subconscious and conscious opinions of people in relation to summarisation tools. Phase 2 addressed the remaining four hypotheses (H2-H5), as follows:

- A: Evaluating the success of human- and computer-generated summaries for helping people understand a topic.
- B: Understanding the effect that size of summary (number of sentences) has on their usefulness.
- C: Understanding the effect that new information has on the different sized summaries.
- D: Understanding how easy it was for participants to tell if a summary is human- or computer- generated.
- E: Comparing both expectations and reactions to the quality of human- and computer-generated summaries.
- F: Further understanding if and how automatic summarisation tools will be used

#### **Methodology**

15 new participants, 7 male and 8 female, were recruited from the same university, using a bulk email to staff and students. 8 participants were undergraduate and postgraduate students from a range of disciplines, and 7 were staff from a range of positions including administration, teaching, and research staff. Participants ranged between 18 and 60 years old and all 15 described

themselves as frequent users of Google. Again, this sample avoided computer scientists who may be affected by their knowledge of algorithms and text analysis.

Phase 2 involved three main tasks. In the task 1, participants rated the quality of both human and computer summaries. Size of summary, topic, and method of creation were all counterbalanced and covered equally by the end of the study. In task 2, participants were asked to decide the creator of human- and computer-generated summaries. Again, exposure to summaries of different size, topic, and method of creation was counterbalanced. In task 3, participants rated five incrementally sized human-generated summaries, as well as rating the new information in them. Participants repeated this step for five incrementally sized computer-generated summaries. The order of seeing human- and computer-generated summary sets was counterbalanced. At the conclusion of phase 2, all participants were again presented with a short survey and debriefing interview, which collected additional qualitative data about their Likert-scale ratings and again the situations where automatic summarisation tools could be practical within their life.

All human-generated summaries taken from Phase 1 were quality checked before being included in Phase 2. Participants were not constrained by time in these tasks. After completing these three steps in Phase 2, the system presented a short questionnaire. The questionnaire asked participants to rate the general quality of human- and computer-generated summaries on a 9-point Likert scale, as well as speculations about their use in search engines. The session was concluded with a short debriefing interview.

## Results

### A - Quality of Human- and Computer-generated Summaries.

Across the 3 tasks, participants rated 135 human-generated and 135 computer-generated summaries for how successfully they summarised their associated topic. Human summaries were rated, on average, as 6.49 out of 9, and computer-generated summaries were given 3.97. Participants gave higher ratings to those summaries that they felt outlined the important information whilst leaving out the unimportant information that bulks out a piece of text. A Mann-Whitney U test revealed that human-generated summaries were significantly more successful than computer-generated summaries ( $U=3525.5$ ,  $p<0.0001$ ). Consequently, H3 was accepted, concluding that human summaries outperform automatic summaries using current state of the art technology.

When rating summaries in task 3, participants were able to see and compare different sized summaries of the same text. A Mann-Whitney U test revealed that human-generated summaries were rated significantly lower in quality when seen in comparison ( $U=1730.3$ ,  $p<0.05$ ). Computer-generated summaries were not rated significantly differently, when shown in comparison to different sized

summaries. This latter finding perhaps indicates that the computer-generated summaries were more consistent or robust, despite being lower in quality.

### B - Relevance of Length on the Quality of Summaries

We hypothesized that the quality of both human-generated summaries (H4) and computer-generated summaries (H5) would increase as more sentences get added to the summary. When asked on the ideal length of a summary, participants felt that around four sentences was a good base to work on but also felt that the length should correspond to the amount of relevant information that is available. [FIG] shows the average summary ratings for both human- and computer-generated summaries for each of the 5 sizes.

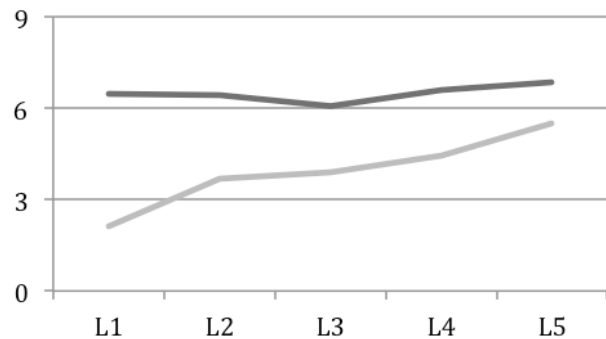


Figure 2 –The effect of summary size on the quality of human-generated (dark) and computer-generated (light) summaries.

While human-generated summaries did not vary significantly in quality across the five sizes (reject H4), H5 was accepted. Computer-generated summaries improved significantly ( $F(4)=20.24$ ,  $p<0.0001$ ) as size of summary increased. A post-hoc Tukey test revealed that the significance was primarily between L1 and L2 ( $p<0.01$ ), but that L5 summaries were rated significantly higher than L1-L3 (all  $p<0.01$ ).

We can conclude from [FIG] that computer-generated summaries approached the quality of human-generated summaries as size increased. Further work is needed to determine when they become comparable, however, as even at Level 5, a Mann-Whitney U test showed that human-generated summaries were rated higher than computer-generated summaries ( $U=214.5$ ,  $p<0.005$ ).

### C – The impact of new information in larger summaries

In task 3 of Phase 2, participants rated the quality of the new information between summaries of increasing size. Rather than measure the quality of new information alone, which would ideally be strong at every increment (no trend was seen in the quality of new information at each level), we wanted to measure the likelihood that new information would affect the quality of the summary. To do this, we subtracted the difference in quality between 2 sizes of summary from the rating given to the new information.

$$Effect_n = New Info_n - (Quality_n - Quality_{n-1})$$

Using this formula, an effect rating of zero would mean that the useful new information was the sole cause of the improved score. Conversely a higher effect rating means the new information had a significantly weaker effect.

Difference in quality and new useful information		L2	L3	L4	L5
Human	Mean	4.13	6.27	4.80	5.07
	Std Dev	2.00	3.17	2.14	2.43
Computer	Mean	2.73	3.67	3.47	3.93
	Std Dev	1.88	1.81	1.93	1.81

**Table 3: Table showing the effect the new information has on the summary quality for lengths from two-five sentences.**

The scores shown in [TABLE] indicate that the new information had a more substantial impact on computer-generated summaries. The poorer scores for human-summaries indicate that they were more likely to include higher-quality detail in every sentence. From these results, we have established that the success of computer-generated summaries is dependent on the new useful information that is contained within them. By integrating entirely information into the summaries, people are more likely to rate the use of the tool as a success.

*D - Predicting the author of a summary*

Task 2 asked participants to decide whether a human or a computer had generated each summary. Despite participants having consistently given human-generated summaries higher scores for quality, participants incorrectly identified computer-generated summaries 40% of the time. While it was easier to identify human-generated summaries (correct 75% of the time), participants found it substantially harder ( $X^2=3.08, p=0.079$ ) to tell whether a computer had generated a summary or not. As the ability to identify computer-generated summaries tended towards 50/50, we chose to reject H2, concluding that participants often struggled to identify computer-generated summaries even though they were not quite as accurate.

To understand why H2 failed, we looked at the reasoning participants gave in regards to their choice of who wrote each summary. The reasoning that people gave for saying that computers wrote summaries ranged from lack of coherence, absence of emotion, impersonal and disjointed text where as they felt human summaries flowed and was very personal and emotive.

*E - Expected/Actual Quality of Summaries Overall*

At the end of Phase 2, we asked participants to reflect, using a 9-point Likert scale, on the ability of humans and computers to generate summaries. These scores, along with the expectations gathered in Phase 1, are shown in [TABLE]

Phase	Question	Mean (std)
1	What was the overall quality of your summaries?	5.50 (2.13)
1	How well do you think computers could generate these summaries?	5.30 (1.63)
2	How well do you think humans can generated summaries?	7.27 (1.06)
2	How well do you think computers can generate automatic summaries?	4.67 (1.49)

**Table 4: The average quality ratings given human and computer-generated summaries in phase 1 and phase 2.**

Having seen and rated both human- and computer-generated summaries (although not knowing which was which), participant thought that humans produced significantly better summaries ( $U=17.5, p<0.0001$ ). Prior to having direct experience with them, participants had higher expectations about computer-generated summaries (5.3 compared to 4.67), although this difference was not significantly different. After Phase 2, however, participants had significantly higher ( $U=111.5, p<0.05$ ) expectations about the quality of human-generated summaries than after having written their own summaries (7.27 compared to 5.5). There are two possible explanations for this: 1) that participants were more conservative about their own summaries in Phase 1, or 2) that experience with computer-generated summaries in Phase 2 gave participants better grounds for making judgements.

*F - Situations where Automatic Summarisation Is Useful*

Just like phase 1, in phase 2 we asked all participants the situations that an automatic summarisation tool might be useful to have. The results in this phase will allow us to see if the situations are similar even though we are no longer dealing with expectations rather those that have had experience with the tool.

The situations that were presented once again conformed to an underlying motivation. Participants felt that situations that an automatic summarisation tool would be useful were those that were lengthy and time consuming to read where participants would want to get a gist of what the text was saying. Situations presented included web pages (2 mentions), research papers (3 mentions) and report reading (2 mentions). However, unlike phase 1, participants were a bit more conscious citing trust in the reliability of the tool as well as not using it for precise knowledge as the reason why it would not be applicable in certain situations.

## DISCUSSION

The aim of this study has been to examine expectations and actual reactions to automatically generated summaries, using human-generated summaries to provide context. Participants in Phase 1, who had not experienced any computer-generated summaries, expected that computers would be able to produce summaries approximately as well as humans (computer 5.30 out of 9; humans 5.50 out of 9). Despite rating computer-generated summaries as an average of 3.69, or relatively unsuccessful at summarising a topic, participants' general expectation about the quality computer-generated summaries, at the end of Phase 2, was 4.67. This score is not significantly different from those in Phase 1 who had not actually seen or rated any computer-generated summaries. The average score for 5-sentence computer-generated summaries, however, was 5.48, which exceeds both expectations. Further, two thirds of the incorrectly judged computer-generated summaries were four or five sentences long.

Considering the difficulty participants had in correctly identifying longer computer-generated summaries, these findings highlight an important difference between expectations and reality. Should search services, for example, include stronger, longer, computer-generated summaries, many users may not realise that the summaries are automatically generated. Even if they did, many would expect them to be comparable to human-generated summaries anyway.

There are several key findings that can be extracted from the two phases of the study:

- With no experience of computer-generated summaries, participants believed that human- and computer-generated summaries would be approximately the same.
- Without knowing which was which, participants rated human-generated summaries higher than computer-generated summaries.
- Participants believed that 40% of computer-generated summaries were written by humans.
- Computer-generated summaries improved significantly with size, reducing the gap with human-generated summaries.
- Larger computer-generated summaries exceeded the expectations of participants.
- The better human-generated summaries were rated lower than the expectations of participants.
- Participants rated human-generated summaries as lower quality, when seen in comparison to other summaries.
- Participants both consciously (with subjective ratings) and sub-consciously (through rating anonymous summaries) rated human-based summaries as better than computer-based summaries.

## Implications

These results seem to indicate that there is great opportunity for information systems to include automatic summarisation to support searchers in the later stages of the IS process: analysing results and identifying valuable sources of information.

Further, the findings above have several implications for design in relation to the potential use of automatic summaries in information systems. The implications for design we have established are:

- Careful consideration needs to be taken in how and where summaries are presented. Showing summaries in comparison to others, such as for each of 10 search results, may reduce their utility.
- Experience with poor summaries may reduce expectations about the utility of the automatically-generated summaries and the search system as a whole.

As well as implications for design, there are also implications for the algorithms being used in text summarisation tools, which include:

- When developing summaries, algorithms should concentrate on key facts and figures that are relevant and avoid text which does not directly relate to the query.
- Summaries should have a sufficient length (5 or more sentences for the technology used here) to provide useful support for sense-making and search.

Beyond implications for design and algorithms, our results would indicate that Human generated summaries should be the main comparison data source when trying to improve summarisation algorithms.

## Future Research

Techniques for automatic summarisation have been maturing over the last ten years so that they are becoming increasingly available for use in everyday information systems. This research, however, has only just begun the process of understanding the utility and usability of automatic summarisation. This research has identified several novel findings relating to peoples expectations about, and reactions to automatic summarisation. Much more work is needed to understand how useful they are for achieving actual learning and sense-making tasks. Task-based lab studies could be used to compare systems that do and do not utilise automatic summarisation during search. Summaries could be compared to other document surrogates like the text-snippets shown with results in most search engines. Beyond analysing utility in artificial settings, longitudinal approaches can be used to study their use in the real world.

## CONCLUSION

While the majority of research into automatically generated summaries has focused on comparative benchmarks with gold-standards, summarisation algorithms have matured and are becoming increasingly available for use in everyday systems. This research has taken the first few steps in exploring their eventual utility and usability. To begin this process, we performed a two-phase study that explored both expectations and actual reactions to automatically generated summaries.

Our research has contributed several novel findings, including that although automatic summaries are still not as strong as human-generated summaries, longer and stronger computer-generated summaries out performed the expectation. Further, participants believed that 40% of computer-generated summaries were generated by humans. These results indicate that both the quality of summaries, created from openly available summarisation services, [5]combined with the expectations of every-day computer users, has created a great opportunity to include summaries in information systems.

## REFERENCES

1. Microsoft. *Technet - Changes in Office 2010*. 2010 [cited 2011 1st September]; Available from: <http://technet.microsoft.com/en-us/library/cc179199.aspx>.
2. Kao, A. and S.R. Poteet, *Natural Language Processing and Text Mining*2007: Springer.
3. Baeza-Yates, R. and B. Ribeiro-Nero, *Modern Information Retrieval*1999: Pearson Education Ltd.
4. Marchionini, G., *Information Seeking in Electronic Environments*1995: Cambridge University Press.
5. Geroimenko, V. and C. Chen, *Visualising the Semantic Web: XML-Based Internet and Information Visualisation*2006: Birkhauser.
6. Witten, I., A. Moffatt, and T. Bell, *Managing Gigabytes: Compressing and Indexing Documents and Images*1999: Morgan Kaufmann.
7. Brin, S. and L. Page, *The Anatomy of a Large Scale Hypertextual Web Search Engine*, in *Seventh International World Wide Web (WWW) Conference*1998: Brisbane, Australia.
8. Google. *Google Instant - About*. 2010 [cited 2011 12th August]; Available from: <http://www.google.com/instant/>.
9. NIST, *TREC: Experiment and Evaluation in Information Retrieval*2005: MIT Press.
10. Hu, V., et al., *Effects of Search Success on Search Engine Re-Usage*, in *CIKM2011*: Glasgow.
11. Sicilia, M.A. and D.L. Miltiadis, *Guidelines for Web Search Engines - From Searching and Filtering to Interface*, in *Metadata and Semantics*2008, Springer. p. 390-391.
12. Rutheven, I., M. Lalmas, and K.V. Rijsbergen, *Incorporating User Search Behaviour into Relevance Feedback*. American Society for Information Science and Technology, 2003. 54(6).
13. Mashable. *Google+ Posts Now Appear in Google Search Results*. 2011 [cited 2011 1st September]; Available from: <http://mashable.com/2011/08/12/google-plus-social-search/>.
14. Herlocker, J., et al., *Evaluating Collaborative Filtering Recommender Systems*. ACM Transactions on Information Systems, 2004. 22(1): p. 5-53.
15. Mani, I. and M.T. Maybury, *Advances in Automatic Text Summarization*1999: MIT Press.
16. Zhou, L., et al. *ParaEval: Using Paraphrases to Evaluate Summaries Automatically*. in *HLT-NAACL*. 2006.
17. Hovy, E. and L. Chin-Yew. *Automated Text Summarization and the Summarist System*. in *Proceedings of the TIPSTER Text Program*. 1998.
18. Mani, I., *Automatic Summarization*2001: John Benjamins Publishing Company.
19. President, S. and B. Dorr, *Text Summarization Evaluation: Correlating Human Performance on an Extrinsic Task with Automatic Intrinsic Evaluation*. 2006.
20. Nenkova, A., *Summarization Evaluation for Text and Speech: Issues and Approaches*. 2006.
21. Lin, C.Y. *ROUGE: Recall Orientated Understudy for Gisting Evaluation*. 2007 [cited 2011 5th September]; Available from: <http://berouge.com/default.aspx>.
22. Maynard, D., et al. *Using a Text Engineering Framework to Build an Extendable and Portable IE-Based Summarisation System*. in *Workshop on Automatic Summarization*. 2002.
23. Krayewski, K. *Cpedia: A Good Idea, in Theory*. 2010.
24. Duan, M. *Cuil Search Engine goes out Quietly*. Silicon Valley/San Jose Business Journals, 2010.
25. Medelyan, O., et al., *Mining Meaning from Wikipedia*. International Journal of Human Computer Studies, 2009. 67(9).

**The columns on the last page should be of approximately equal length.**

**Remove these two lines from your final version.**