Constructing Query-biased Summaries: a Comparison of Human and System Generated Snippets

Lorena Leal Bando
School of Computer Science and IT
RMIT University
Melbourne, Australia
lorena.lealbando@rmit.edu.au

Falk Scholer
School of Computer Science and IT
RMIT University
Melbourne, Australia
falk.scholer@rmit.edu.au

Andrew Turpin
Dept. of Computer Science Software Engineering
The University of Melbourne
Melbourne, Australia
aturpin@unimelb.edu.au

ABSTRACT

Modern search engines display a summary for each ranked document that is returned in response to a query. These summaries typically include a snippet – a collection of text fragments from the underlying document – that has some relation to the query that is being answered.

In this study we investigate how 10 humans construct snippets: participants first generate their own natural language snippet, and then separately extract a snippet by choosing text fragments, for four queries related to two documents. By mapping their generated snippets back to text fragments in the source document using eye tracking data, we observe that participants extract these same pieces of text around 73% of the time when creating their extractive snippets. In comparison, we notice that automated approaches for extracting snippets only use these same fragments 22% of the time. However, when the automated methods are evaluated using a position-independent bag-of-words approach, as typically used in the research literature for evaluating snippets, they appear to be much more competitive, with only a 24 point difference in coverage, compared to the human extractive snippets. While there is a 51 point difference when word position is taken into account.

In addition to demonstrating this large scope for improvement in snippet generation algorithms with our novel methodology, we also offer a series of observations on the behaviour of participants as they constructed their snippets.

Categories and Subject Descriptors

H [**Information Systems**]; H.3.3 [**Information Search and Retrieval**]: Search process

General Terms

Experimentation

IIiX 2010, August 18–21, 2010, New Brunswick, New Jersey, USA.
Copyright 2010 ACM 978-1-4503-0247-0/10/08 ...$10.00.

Keywords

User Studies, Human Summarization, Automated Snippet Generation, Eye Tracking

1. INTRODUCTION

Information retrieval systems help users to find relevant information sources in large, heterogeneous data collections such as the Web. In particular, search engines take a set of keywords provided by the user as a query, and return a ranked list of supposedly relevant documents. For each document in the list, the title, a short summary (often referred to as a snippet or caption), and the URL of the discovered source is displayed.

In this paper we are particularly interested in studying the summary component that is returned by search engines, which is short and biased towards query words. We will use the term *snippet* to explicitly refer to these types of summaries. Currently, Web search engines construct snippets from two or three sentences extracted from the document that have a close relationship with query terms [15]. While extensive research has been conducted into devising and evaluating automatic summarization systems with and without human participation, there are very few papers exploring the human construction of snippets. Typically, in summarization studies, subjects are given instruction to "summarise this document", with no specific information need in mind. The set of human generated summaries then becomes the *gold standard* against which automatic methods are judged.

Our research is motivated by this apparent lack of investigation into how humans construct snippets: a short summary of a single document when there is a specific information need or query. By studying this in detail, we aim to inform the development of automatic snippet generation algorithms.

Towards this goal, this paper presents results from a small user study that explores human snippet generation. Participants were presented with a task, information need and a document, and then asked to firstly create a *generative snippet*, where they were free to write anything they chose. Secondly, the participants were asked to create an *extractive snippet*, where the snippet was constructed from document parts (words, phrases, or sentences) selected from the source document. Eye tracking techniques were used to monitor their reading patterns while they performed study tasks.

Each generative snippet constructed by a participant is an expression of their preferred summary of that particular document and information need. Therefore, the generative snippets are used as the gold standard snippets against which others are judged, and that automatic snippet generation algorithms should aspire to produce such snippets. For each document and information need, participants also create an extractive snippet. By its nature, this snippet is more constrained than the generative version, but can be assumed to be the ideal extractive summary of the document, from the view of the creator.

We test three main hypotheses in this research.

H1 From a users' perspective, a good extractive snippet is constructed from the same text fragments that they read when building their generative snippet.

H2 Current algorithms for building extractive snippets do not extract the ideal fragments referred to in H1.

H3 System performance is overestimated when using a bag-of-words evaluation scheme, rather than taking text structure into account.

The experimental framework we used in order to gather data to test these hypotheses is detailed in Section 3. Section 4 and 5 report the relevant results and a series of observations about human behaviour on the snippet generation task.

We find that around 75% of text fragments used by participants in their extractive snippets were from places in the documents that were used as source material for their generative snippets, confirming H1. The three automated snippet generation methods that we tested used only around 22% of these fragments, significantly less than the human participants for three out of four cases, supporting H2. Comparing human- and system-generated snippets using a simpler bag-of-words based metric shows significantly different trends; generally, the performance of the automated methods appears higher, and the human based snippets are rated as less effective, supporting H3.

2. RELATED WORK

The process a human follows to produce a summary has been widely investigated in the field of psychology (see Hidi and Anderson [5] for an overview). Such studies in this field examine the cognitive processes underlying the construction of a summary, given one or more source documents. While different studies have proposed varying numbers of specific macrorules that are employed in the creation of summaries [14], these can be viewed as three broad processes: first, a *selection* process governs the evaluation of source content, with certain sections being selected for inclusion, while others are deleted. Second, a *condensation* process involves the substitution of some source material, either with more general higher level ideas, or with more specific lower level concepts. Third, a *transformation* process leads to the integration and combination of ideas from the source [5]. Typical tasks carried out by participants in psychological experiments involve writing short summaries, in response to general instructions such as to "write a good summary of the text" [2, 20].

Information retrieval systems use summaries to guide users to resources that are relevant to an underlying information need. Rather than showing a general summary of a resource, modern search engines display tailored snippets, aiming to display that part of the resource that is most likely to be descriptive in relation to the specific information need. Studies in the information retrieval domain [15, 19] and industry practice suggest that query-biased snippets may have more utility to users of search engines than generic summaries. However, none of the psychological studies that we surveyed provide a targeted information need as an underlying motivation for the summarization process, or include a specific query that is to be answered.

Snippets displayed by search engines tend to be short, usually in the order of around 50 words. The summaries that have been studied in previous work, on the other hand, are often much longer. We were unable to locate any study that has investigated the human behavior involved in producing a search engine-style snippet. The Text Analysis Conference (TAC) has included a task where multiple documents are summarized relative to an information need [3], but again, this differs from a snippet generation task where a single document is the object of summarization.

A third feature of the snippet generation problem is introduced by modern computational limitations. As discussed previously, there are at least three cognitive processes that are employed by humans when creating summaries: selection, condensation, and transformation. However, the construction of computer algorithms and systems to emulate all three of these processes at a level comparable to that displayed by humans is beyond current capabilities. Moreover, while natural language technologies are making inroads on the more complex condensation and transformation processes, in the context of search engines summaries need to be generated in microseconds, so that search results can be displayed to users in real time [18]. As a consequence of these factors, current snippet construction algorithms are *extractive*: that is, snippets are created by concatenating pieces of text from the document, rather than attempting to generate new prose. This approach corresponds only to the selection process of the human cognitive model, removing the computationally more complex condensation and transformation processes. In this paper, we examine and compare how humans construct both extractive and generative summaries.

Automatic snippet extraction was first studied by Luhn, who proposed an algorithm using sentences as the minimal unit of extraction. Individual sentences are weighted according to the occurrence of significant words – frequently occurring terms other than stopwords. The top scoring sentences are then chosen for inclusion in the summary [11]. Variation on this approach have included scoring sentences by their position in text and vocabulary transitions [4]. These techniques are query independent; based on a fixed set of parameters, a unique snippet is created for a document, no matter what the underlying information need of the user might be.

Early search engine approaches for snippet generation included showing static summaries. For example, such a snippet might be constructed from the first n sentences that occur in a document (such an approach was used by early versions of the AltaVista search engine), or text located in the metadata tag. As with the Luhn approach, using such static summaries is efficient, since they can be pre-stored and are cheap to fetch at retrieval time. However, these generic excerpts are limited when users have a diverse range of information needs for which documents may be retrieved.

This led to the introduction of *query-biased* snippets, based on the idea that the most relevant fragment of information about a document is likely to differ depending on what the searcher is looking for. Tombros and Sanderson [15] compared query biased snippets with static snippets, and demonstrated that users were better able to determine the relevance of an underlying document when presented with query-biased representations. Several alternative approaches for snippet generation have been proposed in the literature, for example based on the use of document titles [7], or using past queries that have successfully retrieved a document to generate a snippet description [13]. The algorithmic costs of snippet generation in search engines has also been studied [16, 18].

Studying humans while they summarize text may provide valuable insights into the summarization process, inform snippet generation algorithms, and suggest how the quality of different snippets can be evaluated. Jing et al. have investigated agreement among humans in what they perceive as important parts of a document that should be included in an excerpt [6]. Summaries made by subjects have also been used to evaluate automatic summarisation systems. For example, Zechner compared an extractive summary generation approach by comparing the system output with sentences that have been judged by humans to be highly representative of document content [21].

Lin and Hovy proposed the widely-used ROUGE metric, which compares term co-occurrence counts between system generated summaries and a set of human generated reference summaries [9, 10]. Variations of this metric take context into account by counting occurrences of *n-grams* rather than single words, or by considering longest common sequences between the system and reference summaries (ROUGE-L). In this paper, we calculate the coverage between extractive and generative snippets, to investigate the effectiveness of both human and system generated query-biased snippets.

3. EXPERIMENT DESIGN

The study required participants to read a document and an information need, then construct a generative and extractive snippet for that need. They were then shown a second information need for the same document, and required to construct new generative and extractive snippets. Each participant constructed both types of snippets, for two different information needs, and for each of two documents.

3.1 Material and Task

We used two documents (which we call document 1 and 2), each with two queries (query A and B), in this study. The documents and queries were chosen from the *LA Times* subset of the TREC newswire collection according to the following criteria.

- The document should be assessed by TREC assessors as relevant for both queries. This also has the side effect of ensuring that the document has been returned by a modern search engine for the queries.
- The documents should be no longer than 1000 words.
- The content of the document should be readily understandable by an English speaking undergraduate student.

Table 1: Details of the two documents used.

Feature	Doc. 1	Doc. 2
Paragraphs	6	6
Sentences	30	23
Number of words	555	453
Av. sentence length (words)	18.9	18.6
Short sentences ($\leq$ 14 words)	13	11
Long sentences ($\geq$ 29 words)	3	2
Flesch Index	66.0/100	56.7/100
Kincaid (grade level)	8.7	10
Fog Index (years of education)	11.5	13.4

The main features of the selected documents are described in Table 1. Documents had similar length and readability scores. The scores were computed using the Unix command `style`, which also counts the number of long and short sentences as defined in the table. The Flesch index (0 hard, 100 easy) indicates that both documents are easily understandable by 13 to 15 year old students. According to the Kincaid index, documents one and two are suitable for 9th and 10th graders, while the Fog index indicates that undergraduates should not have difficulty with the documents. The two topic queries for each of the documents were presented to participants as a simulated task, following Borlund's approach [1]. The specific queries are shown in italics in Table 2. We focus on informational queries because snippets are more useful for informational rather than navigational information needs [17].

3.2 Subjects and Experimental Procedure

Ten students from RMIT University, aged between 18 and 40, took part in this experiment. Most of them were enrolled in a Computer Science program, and self-reported wide experience using search engines. Eight volunteers had a bilingual or multilingual background with English as their second language, whereas two were native speakers (participants 9 and 10). Only one volunteer reported having any prior knowledge about the topics involved in the study (participant 8). To record eye movements, we used a Tobii T60 eye tracker, which is integrated with a 17" TFT display, a resolution of 1280x1024 pixels and a frequency of 60Hz.

General instructions and a training exercise were presented to maintain uniformity between subjects. Before the study was carried out, volunteers were informed that they could perform the experiment with no restrictions on time.

Firstly, subjects were asked to read the simulated work task (without the requests), and then to read the associated document. Once they were satisfied that they had read the document, the interface displayed one of the specific information needs (or requests). The order of documents and information needs was randomized and balanced across participants. The subject was then asked to write a short summary (no more than 400 characters) related to the information need. The document and query remained displayed during this part of the study in case volunteers needed to consult certain details; however the interface was designed to not allow participants to copy and paste. They were reminded to employ their own words and summarisation strategies, and not to be concerned about grammatical mistakes. Participants stopped writing when they considered their summary was sufficient to complete the task. If

Table 2: Simulated work tasks and requests for document 1 and 2, respectively.

Simulated work task 1. Your friend has become a member of an animal protection group and will present a talk about endangered species and wildlife next week. Your friend has been consulting several sources and has asked to you to help summarise a document according to specific questions that could arise from the audience. **Request A**. One question that your friend needs to prepare for is: *Why are pandas considered to be endangered?* Identify *their habitat* and explain *what threatens them.* **Request B**. The document could be useful to prepare for the questions: *What efforts have been made to prevent the demise of pandas? Which countries are making such efforts?*
Simulated work task 2. As a member of a scientific group, you will undertake an expedition to Antarctica. You have been asked to research prior explorations and the role of krill in Antarctica. **Request A**. Your team leader is interested in the following information: Identify *current or planned systematic explorations and scientific investigations of Antarctica.* **Request B**. Your team leader is interested in the following information: *What are krill? Why are they important to Antarctica?*

the summary exceeded 400 characters in length, subjects had to modify the summary content and reduce its length. This summary is referred to as the *generative snippet* for each particular participant-document-query combination.

Once participants finished with their generative snippet, the interface provided the same request, but they constructed a summary by selecting parts of the document. We refer to whole sentences, phrases, or words that participants extracted as *parts* of the document. After they selected any part of text, that selection could be copied into a summary area. In this area, participants could customise their summary by reordering or deleting previously extracted elements. They were allowed to select from one to eight document parts to create their summary. The final summary they constructed is referred to as their *extractive snippet.* Once participants completed both generative and extractive snippets for a query, they performed the same tasks for a second query on same document. They then repeated the whole procedure for the second document.

3.3 Mapping Generative Snippets to Document Fragments

To investigate our first hypothesis – that for a participant, their ideal extractive snippet when given a particular document and information need is constructed from the same text fragments that they read when building their generative snippet – for each generative snippet we had to identify the set of text fragments in a document that had been used to construct that snippet. Accordingly, the eye tracking recording of the construction of each generative snippet was manually examined by the first author. The data was processed in two phases. Firstly, the fragment of text that was read immediately prior to the participant typing a section of their snippet was identified. Secondly, the first author made a subjective judgement on the content of the identified text fragment and the section of the generative snippet to ensure that the selection was feasible. When the last read fragment was unrelated to the typed text, it was discarded.

Figure 1 shows an example of the eye tracking data that was available to us. The position of gaze was used to define the window of time that was considered "reading immediately prior" to the participant typing their text, as follows. Beginning with the single fixation immediately prior to the typing, the window of time was extended backwards while fixations were within the same paragraph of text. That is, no large eye movements away from the text region were made. Within this window, the heat-map of gaze in the participant's recording generated by the Tobii was used as an indicator of where "reading" took place. Fragments of text within the region that were covered by contiguous fixations, typical of a reading pattern, were chosen when the recording showed a fixation path from left to right, or vice versa in case of re-readings [12]. Figure 1 illustrates an example of fixations and the original generative text.

In some instances it was not obvious which part of the text had been read immediately prior to the construction of the generative snippet. For example, perhaps participants relied on memory of the text to build their snippet, rather than reading. Hence it was possible for some portions of generative summaries to have no related text fragments from the documents. We report this data in Table 3 in the next section.

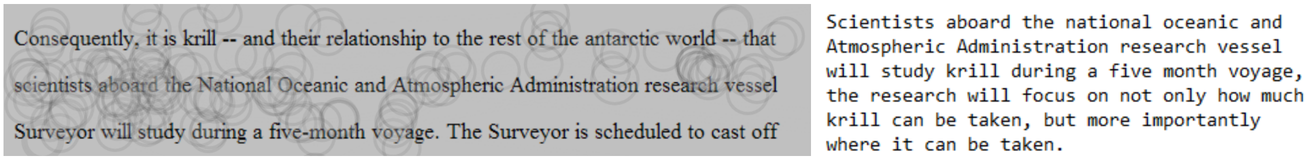
3.4 Automated Methods

In addition to human generated snippets, we also implemented three automated approaches which produced snippets that were 15% of the document size – that is, 5 sentences for Document 1, and 3 sentences for Document 2.

The first method, L, is *Luhn*'s approach, which scores sentences according to their relation with significant words, but takes no account of query terms [11]. The set of significant words is constructed by discarding very frequent terms, such as stopwords, and terms with frequency less than 3. Then, sentences may contain one or several clusters which are composed of significant and non-significant words. A sentence score is defined by w^2/c, where w represents the number of significant words in a cluster and c is the cluster size. If the sentence has more than one cluster, the score is determined by the highest cluster value. Top ranked sentences are extracted and presented in order of appearance in the document.

The second method, Q, includes a *query term* bias, so we used the italicized parts of the requests in Table 2 to form the query term sets. Stopwords were removed from the queries. Each sentence was weighted by s^2/q, where s is the number of query terms in a sentence and q is the total number of query terms [15]. This approach selects highly scored sentences, and returns them ordered by decreasing score.

The third method, P, adds a constant *positional* bias [18]



Consequently, it is krill -- and their relationship to the rest of the antarctic world -- that scientists aboard the National Oceanic and Atmospheric Administration research vessel Surveyor will study during a five-month voyage. The Surveyor is scheduled to cast off

(a)

Scientists aboard the national oceanic and Atmospheric Administration research vessel will study krill during a five month voyage, the research will focus on not only how much krill can be taken, but more importantly where it can be taken.

(b)

Figure 1: An example of the eye tracking data that was collected as part of this study. The gaze plot of participant's fixations prior to writing G is displayed in first panel. The gazeplot of $G_{10}^{2,A}$ was generated by participant 10. The generative sentence is presented in the second panel.

to the L and Q scores

$$p = \begin{cases} 1 & \text{if sentence is the document title or the second} \\ & \text{document sentence} \\ 2 & \text{if sentence is the first document sentence} \\ 0 & \text{otherwise} \end{cases}$$

to get a total score for each sentence. In this method significant terms were obtained as described by Tombros and Sanderson [15]. The top-scoring sentences are returned, and displayed ordered by decreasing score.

4. RESULTS

Analysis of results are divided into two sections. In the first, the inclusion of a word in generative and extractive snippets is studied based on its current position in document. In contrast, the second analysis relies on a typical bag-of-words scheme.

4.1 Position-dependent Analysis

The parts of Document 1 that were used in different types of snippets are shown in Figure 2. Each square represents a word in the source document, while query terms are shown as triangles. The left panel shows the frequency with which selected regions were included in extractive snippets by participants; the middle panel shows the frequency of regions viewed when creating generative snippets; and the right panel shows the regions selected by the three automatic approaches. In each case, darker shading indicates a higher frequency of selection. The figure enables direct comparison of selected regions between the three groups. For example, in the extractive snippets, document parts from paragraphs 2, 5 and 6 were particularly popular. For generative snippets, parts of these same paragraphs were also selected. The automated methods, on the other hand, select no document sections from paragraphs 5 or 6.

We now quantify the coverage between the different types of snippets more formally. Let $E_i^{d,q}$ be the set of words used in the snippet extracted by participant i for query q and document d. Similarly, let $G_i^{d,q}$ be the set of words used in the generative snippet constructed by participant i for query q and document d. There were 4 sentences in the total of 119 generated by all participants that were clearly factually incorrect; that is, they directly contradicted or were not supported by information in the document. These four sentences were deleted from the data set.

For any extractive snippet, we define $e_i^{d,q}$ to be the set of all positions of the words used in $E_i^{d,q}$. For example, the shaded squares in the left panel of Figure 2 shows $\bigcup_{i=1}^{10} e_i^{1,A}$ for Document 1 and Query A, as each word that occurs in some extractive snippet has a frequency greater than zero.

Table 3: Number of words that could be identified in the original document for each generative snippet (g_i) and the number of words used in extractive snippets (e_i). The final three rows represent the three automated methods.

	1,A		1,B		2,A		2,B	
i	$\lvert g_i \rvert$	$\lvert e_i \rvert$	$\lvert g_i \rvert$	$\lvert e_i \rvert$	$\lvert g_i \rvert$	$\lvert e_i \rvert$	$\lvert g_i \rvert$	$\lvert e_i \rvert$
1	33	101	34	99	23	74	35	118
2	21	57	34	59	47	83	24	45
3	11	39	–	–	35	61	6	22
4	42	55	–	–	74	92	51	50
5	–	–	36	28	86	51	10	44
6	9	55	31	110	45	52	–	–
7	17	104	45	71	49	66	63	60
8	73	73	84	54	24	38	47	41
9	57	86	54	55	27	130	32	78
10	52	105	51	138	69	92	63	44
L	–	138	–	138	–	62	–	62
Q	–	126	–	126	–	60	–	63
P	–	131	–	131	–	62	–	62

Note that we know $e_i^{d,q}$ precisely as participants were restricted to cutting and pasting from the document d.

Similarly we define $g_i^{d,q}$ to be the set of all positions in document d of words that were read and occur in the snippet generated by participant i for query q. Hence, the shaded squares in the center panel of Figure 2 shows $\bigcup_{i=1}^{10} g_i^{1,A}$. Note that $g_i^{d,q}$ is not as precise as $e_i^{d,q}$ as it was generated based on eye-tracking data, as described in Section 3.3. As such, some $g_i^{d,q}$ sets may be empty, or very small, as we could not identify the parts of the text from which the generative snippet was drawn. Table 3 shows the size of $g_i^{d,q}$ for the four document-query combinations, and 10 participants. We excluded $g_5^{1,A}$, $g_3^{1,B}$, $g_4^{1,B}$ and $g_6^{2,B}$ from all analysis as we could not identify suitable positions in the documents for these generative snippets.

Our first hypothesis proposes that users would construct their extractive snippets from the same text fragments that they used to construct their generative snippet. That is, $e_i^{d,q} = g_i^{d,q}$ for all d, q, and i, assuming that $g_i^{d,q}$ represents the best possible snippet for document d and query q in the eyes of participant i. Qualitatively, comparing the left two panels of Figure 2 shows that this was the case: generally the areas of the document that participants used in their two snippets were the same. To quantify the proportion of g_i that is contained in e_i, we compute the *coverage* as

$$\frac{|e_i^{d,q} \cap g_i^{d,q}|}{|g_i^{d,q}|} \qquad (1)$$

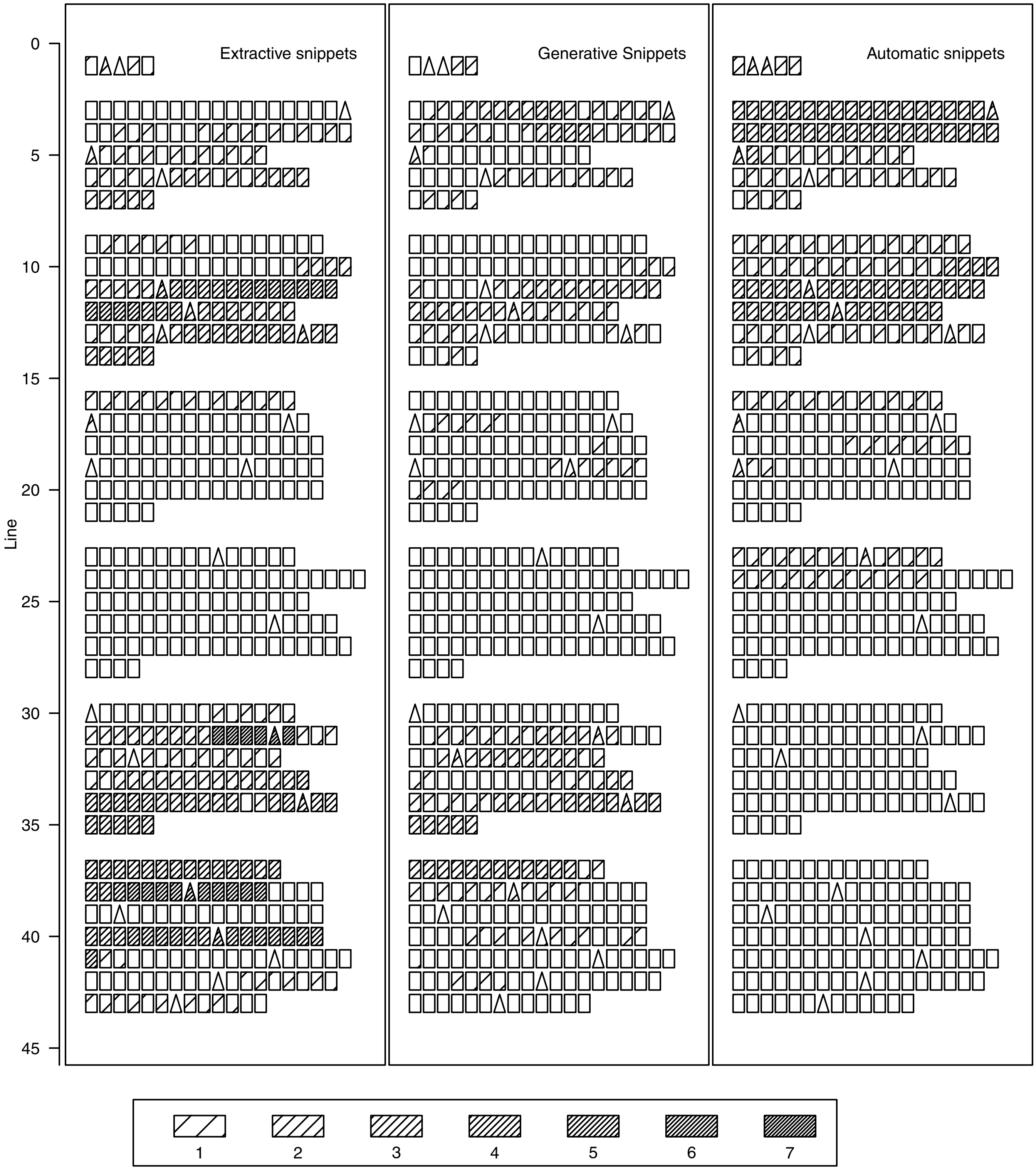
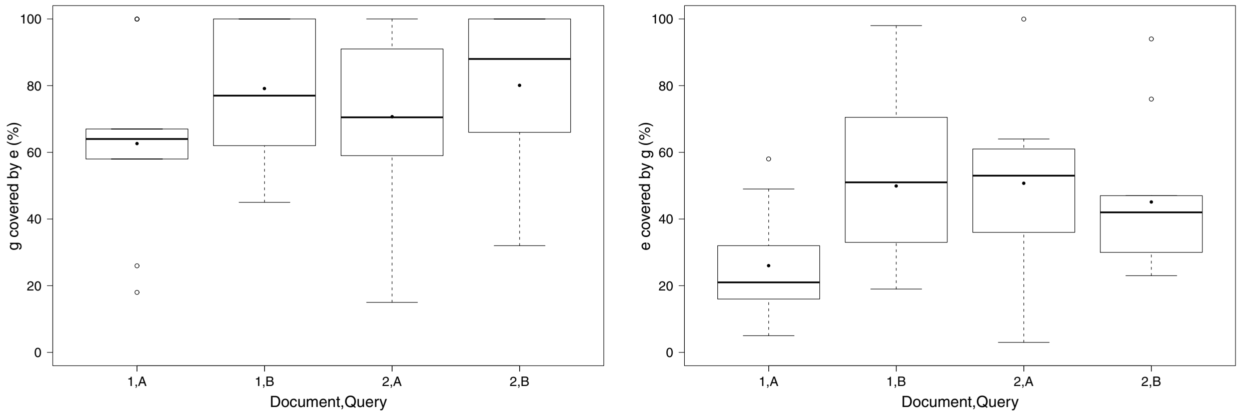


Figure 2: A map of the frequency with which words were used in different snippets for Document 1, Query A. Each word in the document is represented by a square, with the query terms shown as triangles. The level of shading indicates the frequency with which that word was used according to the key at the bottom of the figure. The left panel shows all 10 participant's extractive snippets; the center panel shows all 10 participant's generative snippets; and the right panel the 3 automated methods.



(a)

(b)

Figure 3: (a) Percentage of document text fragments that are used in generative snippets and are also used in extractive snippets (b) the reverse. Boxes indicate 25% and 75% quartiles, the solid line the median, the solid dot the mean, and whiskers extreme values.

for all d, q, and i, with the results shown as percentages in Figure 3(a). There is a box-and-whisker for each document-query combination representing the value of Equation 1 for all 10 participants. The box shows the interquartile range, while the solid bar and circle in the box represent the median and mean, respectively. Whiskers and open circles show extreme values. As can be seen, the proportion of the source for generative snippets that is used in extractive snippets is quite high, ranging from a mean of 63% for Document 1 and Query A, to 80% for Document 2 and Query B. On average over all four document-query combinations, 73% of the text underlying generative snippets is covered by the corresponding extractive snippets.

As a point of comparison, we can consider the coverage of g_i that is likely due to a random selection of words to make up e_i. The chance that a randomly drawn set of $|e|$ words from a document leads to $e \cap g = g$ (100% coverage) is given by the Hypergeometric distribution [8]: Hg(k=|g|, n=|e|, m=|g|, N) where N is the number of words in the document. This probability simplifies to:

$$^{N-|g|}C_{|e|-|g|} / ^{N}C_{|e|} \tag{2}$$

Using typical values from Table 3, the probability that a generative snippet of length 30 is covered by a random extractive snippet of length 50 for a 500 word document is less than 10^{-34}. Even with corrections for multiple comparisons, it is reasonable to conclude that our participants were operating far from randomly.

Interestingly, while on occasion the coverage of g_i by e_i is 100%, generally it is not. That is, there are parts of the document that are used for constructing the generative snippet that are not selected to be part of the extractive snippet. There are several reasons why this might be the case, which we discuss in the final section of the paper.

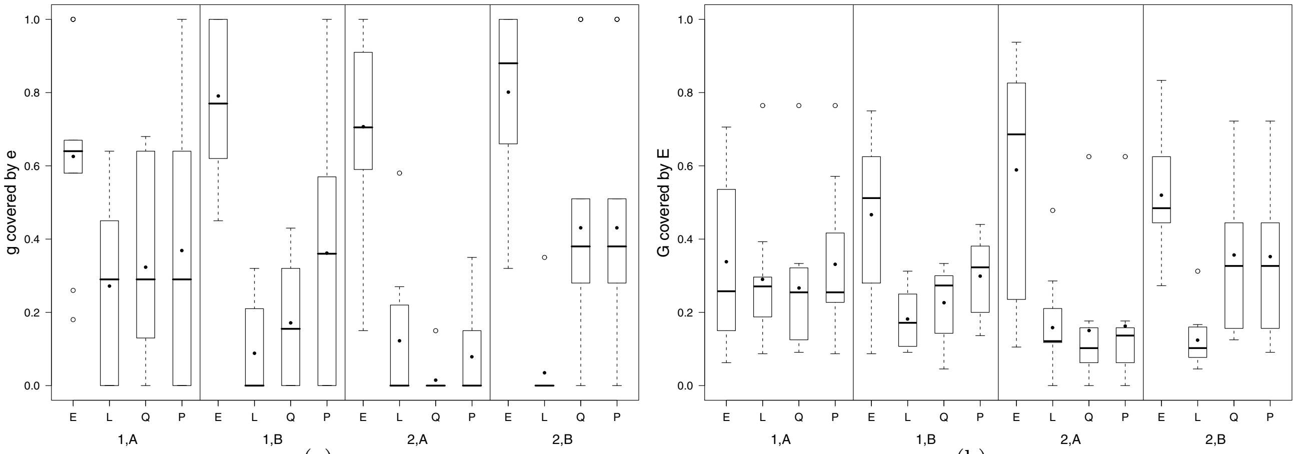
It is of course possible that $e_i^{d,q} = g_i^{d,q}$ because participant i used an extractive technique to construct their generative snippet. That is, they chose not to use more complex language constructions that drew together various parts of the text, and simply copied parts of the text. In such cases, the value of

$$\frac{|e_i^{d,q} \cap g_i^{d,q}|}{|e_i^{d,q}|} \tag{3}$$

should also yield values of 100%. Figure 3(b), however, shows that generally this is not the case: while extractive snippets include the underlying text used in generative snippets, the reverse is not true. It is also apparent that several participants on some d, q pairs did use an extractive technique to construct their generative snippets (as shown by extreme values near 100%), although the number is small (7 data points).

We can compute Equation 1 for the snippets extracted by the three automated systems, and compare these with the extractive snippets created by participants. The results, shown in Figure 4(a), indicate that the automated methods are generally not selecting the same areas of text that were used by the participants to construct their generative snippets. A one-way ANOVA on the snippet type ((E)xtractive, (L)uhn, (Q)uery-biased, and (P)osition-biased) indicates the presence of statistically significant differences in coverage for Document 1 with Query B, and for both queries with Document 2 ($p < 0.001$). A follow-up Tukey Honest Significant Difference test demonstrates that: for Document 1 with Query B, (L)uhn and (Q)uery-biased give significantly lower coverage; for Document 2 with Query A, significantly lower coverage is given by all three automatic systems; and for Document 2 with Query B, only (L)uhn gives significantly lower coverage than (E)xtractive ($p < 0.05$). At a macro level, there is no statistically significant difference in coverage between the two documents (t-test, $p > 0.1$).

At the individual snippet level, the performance of the automated methods is in fact little better than chance. Using the Hypergeometric distribution from above, the mean coverage for a randomly extracted snippet of length 138 words would be 26% for document 1, and 14% for a snippet of length 62 for document 2. This supports our Hypothesis H2: current algorithms for the construction of query-biased snippets do not select the same fragments that users read when constructing their ideal generative snippets.



(a) (b)

Figure 4: Coverage of each participant's generative snippet by the three automatic methods L, Q and P (a) computed using Equation 1 where positions are taken into account $g_i^{d,q}$, and (b) computed using Equation 4 based on bag-of-words data $G_i^{d,q}$.

4.2 Position-independent Analysis

The criteria that an extractive snippet should be drawn from the same locations in the document as a generative snippet is much more restrictive than typical snippet evaluation measures. Usually snippets are evaluated as "bags-of-words", and a good snippet merely must contain the same words as the gold standard. Figure 4(b) shows

$$\frac{|G_i^{d,q} \cap E_i^{d,q}|}{|G_i^{d,q}|}, \tag{4}$$

which no longer takes the position of words from the document into account. Note that stopwords have been removed from E and G. Based on this analysis, the automated methods now generally perform better, showing a higher coverage, while the participants perform worse. In particular, for Document 1 with Query A, shown in the left-most panel of Figure 4(b), there is no statistically significant difference between any snippet types (ANOVA, $p > 0.1$). Each of the other three document-topic pairs show significant effects (ANOVA, $p < 0.001$). For Document 1 with Query B, L and Q are significantly worse than participant-created extractive snippets based on the Tukey HSD test ($p < 0.001$ and $p = 0.002$, respectively), while extractive and P are close to the traditional significance threshold ($p = 0.051$). For Document 2 with Query A (the third panel of the figure), extractive snippets show significantly higher coverage than automatic methods ($p < 0.001$), while differences between the automatic methods are not significant. Finally, for Document 2 with Query B, L snippets perform significantly worse than all other approaches ($p < 0.001$ for L versus extractive, and $p = 0.022$ and 0.025 when comparing L with Q and P, respectively). Again, a t-test shows no significant differences in coverage between the two documents overall ($p = 0.966$), or between the two topics in Document 1 ($p = 0.762$) or Document 2 ($p = 0.192$).

It is clear from the preceding analysis that conclusions about relative levels of coverage between different snippet types are affected strongly depending on whether coverage is measured using the eye-tracking or the bag of words approach.

Table 4: The ratio $|g_i^{q,d}|/|G_i^{q,d}|$ for each of the four d, q combinations and 10 participants.

i	1,A	1,B	2,A	2,B
1	0.825	0.971	0.605	0.636
2	0.777	1.307	1.236	0.648
3	0.239	0.000	0.897	0.139
4	0.792	0.000	1.138	0.796
5	0.000	1.125	2.047	0.212
6	0.257	0.508	1.500	0.000
7	0.377	1.323	1.814	1.750
8	1.074	2.400	1.043	1.305
9	1.118	1.459	0.600	0.780
10	0.867	0.761	1.189	1.536

5. DISCUSSION

In the following lines, we show trends that were observed while participants constructed snippets.

5.1 Generative versus Extractive Snippets

Clearly participants extract pieces of text that are also the source of their generative summaries (Figure 3(a)). In some cases, however, not all of the generative text was chosen to be part of the extractive snippet. That is, the ratio given by Equation 1 was not 100%. There are several possible reasons for this. Firstly, sometimes the extractive snippet is shorter than the generative snippet, which means there is no way the ratio can be 100%. From Table 3 we can see this occurs in 7 instances (for example, $|g_5^{2,A}| > |e_5^{2,A}|$). Secondly, one of the participants exhibited a learning effect, where the first query for each document has a low ratio, but the second is high. As the order of documents and queries was randomized and balanced across all participants, this should not have an overall affect, but it is interesting to observe. Specifically for participant 9:

	$\lvert g_9 \rvert$	$\lvert e_9 \rvert$	Eqn 1	Eqn 3
Document 1, 1st query	57	86	26%	17%
Document 1, 2nd query	54	55	100%	98%
Document 2, 1st query	27	130	15%	3%
Document 2, 2nd query	32	78	100%	41%

Finally it is possible that our extraction of g_i was incomplete due to limitations of the accuracy of the eye-tracking system for a small number of cases. Table 4 shows the ratio of the number of words in $g_i^{d,q}$, the generative snippet constructed by us using eye tracking data, and $G_i^{d,q}$, the actual snippet that the participant typed. Generally our identified snippets are about the same size as the typed snippets, with an overall mean for the ratio of 0.91. This gives us some confidence that we did not introduce length-based artifacts into our analysis. We noticed that it was sometimes difficult to get accurate eye tracking data for participant 3, and that may contribute to their low ratios. That is, we were not able to accurately identify all text regions in the source document for $g_3^{d,q}$.

5.2 Observations on Snippet Creation

In addition to differences in the usage of different words when creating corresponding extractive and generative snippets, the snippets also differed in other dimensions, such as the time required to create them, and the way in which participants chose which information to include.

Overall, participants took more time when creating generative snippets (a mean time of 5 and 3.5 minutes when working with documents A and B, respectively), compared to extractive (with a mean time of only 3.5 and 2 minutes for documents one and two). Note that these results are only suggestive, since generative snippets were always created before extractive ones, so ordering effects may be present and further investigation is required.

While extractive snippets are "correct", in the sense that they use sentences that occur in the source document, generative snippets allow for creativity, so errors may be introduced. From our user study, we collected a total of 119 sentences from generative snippets, which were spelling-corrected and classified based on information they conveyed. A sentence or clause was identified as *factual* when it included correct information in accordance with the document content; *partially correct* if it included information that was not directly traceable to the source content, but was not factually incorrect; *wrong* if it contained incorrect or misunderstood data; or *inferred* when subjects deduced information from own knowledge or document content to produce a new sentence. Over 80% of sentences included in generative summaries were correct and directly traceable to content in the source document, and 11% were partially correct. Only 3.4% of sentences contained incorrect information, and 4.2% of the sentences were produced by deduction.

For current computerised snippet generation approaches, the unit of extraction is usually a sentence [11, 15]. We therefore analysed the units selected by participants when creating an extractive snippet. Less than 50% of the selections made were whole sentences from the source document. In particular, participants avoided extracting prepositions, articles, adverbs and common verb conjugations. These observations can be used to tailor automated snippet generation approaches, maximising the informative content that can be placed on the screen.

It has previously been argued that the title of a document is usually often a good descriptor of document content. For example, early static search engine summarization approaches simply displayed the title and first sentence or two of a target document. In our study, the titles of both documents were closely related with the topic queries. However, the titles were only selected infrequently by participants for inclusion in the extractive snippets: for document one and two, the title was selected by one participant (a visual example of this can be seen in the left-hand panel of Figure 2).

The position within a document at which an informative clause or sentence occurs has been thought be related to the importance of such a fragment. For example, some snippet generating algorithms assume that more informative content occurs towards the start of a document, and therefore include a position bias so that sentences that occur earlier are given a higher probability of being selected [15]. However, our data does not completely support this assertion, with 2 participants referencing information form first paragraph in document one query one, and 1 participant selecting content from the first paragraph for document two query one.

Participants generally chose chunks for inclusion in an extractive snippet by working through a document from top to bottom. More specifically, we examined the relationship between the order in which sentences occurred in the document, and the order in which they appear in the final extracted snippet. Extracted fragments were seldom re-ordered, even though the study interface provided this feature and participants were notified of this. In general 78% of the extractive snippets were constructed using the top to bottom trend. In Document 1, the Pearson correlation coefficient between document and snippet sentence order was $r = 0.574$ ($p < 0.001$) and $r = 0.565$ ($p < 0.001$) for Queries A and B, respectively. For Document 1, although participants generally followed document order, the selections that they chose to include were distributed throughout the document. For Document 2, the order of selection and position in the document were even more strongly correlated, with $r = 0.952$ ($p < 0.001$) and $r = 0.972$ ($p < 0.001$) for Queries A and B, respectively. These observations suggest that automatic snippet generation methods might benefit from preserving the order of extracted sentences to match the order in the original document, even when individual sentence selection schemes might indicate a higher weight for sentences that occur later in the document.

The main difficulty with generative snippets was when participants used two sentences from different informative parts, and generalised their meaning into a single sentence. A similar analysis employing eye tracking data will be conducted to study the relation between the order of written sentences and document content.

6. CONCLUSIONS

In this paper we conducted a pilot user study to examine how humans create generative and extractive snippets, and compared this to current automatic snippet generation approaches. Our results suggest that for humans engaging in a query-biased summarization task, there is a high degree of agreement between those parts of the underlying document that they selected when creating an *extractive snippet* and those parts of the document that they read when creating a *generative snippet*. However, current automatic approaches do not tend to select document parts from these same areas. If the same snippets are compared with a simple bag-of-words approach, human created extractive snippets perform worse, while some of the automated methods perform more strongly. It therefore appears that the performance of the latter is being overestimated by most current snippet evaluation approaches, which do not take the position of the gold standard text in a document into account.

This paper has focused on the snippet generation process, and demonstrated that humans often focus on the same source content whether they are creating generative or extractive snippets. In future work, we intend to investigate differences in the usefulness of the different types of snippets to end-users, compare with more sophisticated summarization approaches, and include more queries and documents.

We also reported a number of observations that will assist in the construction of extractive snippets that more closely approximate the processes employed by humans when engaging in this task. We plan to apply these observations to new snippet generation algorithms that more closely model the selections of humans.

7. ACKNOWLEDGEMENTS

We thank the volunteers who took part in our study. This research is supported by scholarship No. 201099 granted by the Consejo Nacional de Ciencia y Tecnología (CONACyT), and the Australian Research Council.

8. REFERENCES

[1] P. Borlund. Experimental components for the evaluation of interactive information retrieval systems. *Journal of Documentation*, 56:71–90, 2000.

[2] A. Brown and J. Day. Macrorules for summarizing texts: The development of expertise. *Journal of Verbal Learning and Verbal Behavior*, 22(1):1–14, 1983.

[3] H. Dang and K. Owczarzak. Overview of the TAC 2008 Update Summarization Task. In *Proceedings of Text Analysis Conference*, 2008.

[4] H. Edmundson. New methods in automatic extracting. *Journal of the ACM*, 16(2):285, 1969.

[5] S. Hidi and V. Anderson. Producing written summaries: Task demands, cognitive operations, and implications for instruction. *Educational Research*, 56(4):473–493, 1986.

[6] H. Jing, R. Barzilay, K. McKeown, and M. Elhadad. Summarization evaluation methods: Experiments and analysis. In *AAAI Symposium on Intelligent Summarization*, pages 51–59, 1998.

[7] H. Joho, D. Hannah, and J. M. Jose. Emulating query-biased summaries using document titles. In *SIGIR '08: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 709–710, 2008.

[8] K. Krishnamoorthy. *Handbook of Statistical Distributions with Applications*. CRC Press, 2006.

[9] C. Lin and E. Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of HLT-NAACL*, volume 2003, 2003.

[10] C. Lin and F. Och. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of ACL*, pages 606–613, 2004.

[11] H. P. Luhn. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2), 1958.

[12] K. Rayner. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124(3):372–422, 1998.

[13] F. Scholer and H. Williams. Query association for effective retrieval. In *Proceedings of the eleventh international conference on Information and knowledge management*, page 331. ACM, 2002.

[14] C. Sherrard. Teaching students to summarize: Applying textlinguistics. *System*, 17(1):1–11, 1989.

[15] A. Tombros and M. Sanderson. Advantages of query biased summaries in information retrieval. In *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 2–10, 1998.

[16] Y. Tsegay, S. Puglisi, A. Turpin, and J. Zobel. Document compaction for efficient query biased snippet generation. In M. Boughanem, C. Berrut, J. Mothe, and C. Soule-Dupuy, editors, *ECIR 2009 31st European Conference on Information Retrieval*, volume 5478 of *LNCS*, pages 509–520. Springer-Verlag, 2009.

[17] A. Turpin, F. Scholer, B. Billerbeck, and L. Abe. Examining the pseudo-standard web search engine results page. In *Proceedings of the 11th Australasian Document Computing Symposium*, pages 9–16, Brisbane, Australia, 2006.

[18] A. Turpin, Y. Tsegay, D. Hawking, and H. E. Williams. Fast generation of result snippets in web search. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 127–134. ACM, 2007.

[19] R. White, J. Jose, and I. Ruthven. A task-oriented study on the influencing effects of query-biased summarisation in web searching. *Information Processing & Management*, 39(5):707–733, 2003.

[20] P. Winograd. Strategic difficulties in summarizing texts. *Reading Research Quarterly*, 19(4):404–425, 1984.

[21] K. Zechner. Fast generation of abstracts from general domain text corpora by extracting relevant sentences. In *Proceedings of the 16th Conference on Computational Linguistics*, pages 986–989. Association for Computational Linguistics, 1996.