Effective Summarisation for Search Engines

A thesis submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy

Lorena Leal Bando
B.E., M.Sc,
School of Computer Science and Information Technology,
College of Science, Engineering and Health
RMIT University

March 2013

Declaration

I certify that except where due acknowledgement has been made, the work is that of the author alone; the work has not been submitted previously, in whole or in part, to qualify for any other academic award; the content of the thesis is the result of work which has been carried out since the official commencement date of the approved research program; any editorial work, paid or unpaid, carried out by a third party is acknowledged; and, ethics procedures and guidelines have been followed.

Lorena Leal Bando
School of Computer Science and Information Technology
RMIT University
March 2013

Acknowledgments

I am grateful that my PhD journey in Australia has been surrounded by many good friends and colleagues. I thank Dr Falk Scholer and Dr Andrew Turpin for their guidance, advice and support. I deeply appreciate the time you always had for providing me feedback even in your sabbatical or working outside RMIT University. I am also glad that our meetings and discussions were funny.

My family in Mexico is huge so to all of them thanks for cheerful wishes, encouragement and unconditional support. Specially my mum Ofelia, my dad Valeriano and my hilarious brother José. My dad still does not understand how I have spent that many years at Uni. Daniel, my partner and new member of the family, thanks for your caring love and patience during this time.

Thanks to Dr Darnes Vilariño for encouraging me to pursue a doctoral degree. The experience has been rewarding and full of new challenges and goals to achieve.

My PhD could have not been done without the financial support of several funding bodies. Thanks to my sponsor CONACyT, which is the Council of Science and Technology in Mexico, the School of Graduate Research at RMIT University for travelling support to attend to two conferences, and to the Information Storage, Analysis and Retrieval (ISAR) discipline for financing my user studies.

To all past and current fellow students in the Cave (rooms 14.09.03 and 14.09.04) for their help, friendship and sharing student wisdom. To the movie addict group Naimah Yaakob, Mardi Almashor, Vidura Gamini and family for having so much fun outside the research world. My first friend in Australia Nisje Dewi thanks for your warming thoughts and wishes. I am in food debt with Sully Cánovas for feeding me in his asados and pizza afternoons.

I also thank to the members of the writing circle directed by Jennifer Anderson and the Statistics Consultancy team at RMIT University. To the staff members in the ISAR discipline and to CS&IT admin staff, thanks for having a positive attitude with postgrad students. My PhD journey started with the initial guidance of Dr Michael Winikoff, Dr Lin Padgham, Dr Vic Ciesielski and Dr John Thangarajah. Thanks for introducing me to the ISAR troop.

Lastly but not least to my in friends Chelo, Ana Karina, Karina Rosales y Gesuri, who always were asking: “When are you finishing?” So I suppose that after reading this document they can find an answer.

Gracias a todos por su apoyo!

Credits

Portions of the material in this thesis have previously appeared in the following publications:

- L. L. Bando, F. Scholer, and A. Turpin. Constructing query-biased summaries: a comparison of human and system generated snippets. In *Proceedings of the third symposium on Information interaction in context*, pages 195–204, New York, NY, USA, 2010. ACM

- L. L. Bando, F. Scholer, and A. Turpin. Sentence Length Bias in TREC Novelty Track Judgements. In *Proceedings of the 17th Australasian Document Computing Symposium 2012*, pages 55–61. ACM, 2012

Contents

List of Figures

List of Tables

Abstract

Users of information retrieval (IR) systems issue queries to find information in large collections of documents. Nearly all IR systems return answers in the form of a list of results, where each entry typically consists of the title of the underlying document, a link to the document, and a short query-biased summary of a document's content called a snippet. As retrieval systems typically return a mixture of relevant and non-relevant answers, the role of the snippet is to guide users to identify those documents that are likely to be good answers, and to ignore those that are less useful. This thesis focuses on techniques to improve the generation and evaluation of query-biased summaries for informational requests, where users typically need to inspect several documents to fulfil their information needs. We investigate the following issues: how users construct query-biased summaries, and how this compares with current automatic summarisation methods; how query expansion can be applied to sentence-level ranking to improve the quality of query-biased summaries; and, how to evaluate these summarisation approaches using sentence-level relevance data.

First, through an eye tracking study, we investigate the way in which users select information from documents when they are asked to construct a query-biased summary in response to a given search request. Our analysis indicates that user behaviour differs from the assumptions of current state-of-the-art query-biased summarisation approaches. A major cause of difference resulted from vocabulary mismatch, a common IR problem.

This thesis then examines the generation of query-biased summaries as a sentence ranking problem, where existing ranking techniques and evaluation measures for document retrieval are adapted to automatic query-biased summarisation. We study statistical approaches to ranking sentences assisted by query expansion techniques to improve the selection of candidate relevant sentences, and to reduce the vocabulary mismatch observed in the previous study. We employ a Cranfield-based methodology, widely used for measuring document retrieval effectiveness, to quantitatively assess sentence ranking methods based on sentence-

level relevance assessments available in the TREC Novelty track, in line with previous work.

We study two aspects of sentence-level evaluation of this track. First, whether sentences that have been judged based on relevance, as in the TREC Novelty track, can also be considered to be indicative; that is, useful in terms of being part of a query-biased summary and guiding users to make correct document selections. By conducting a crowdsourcing experiment, we find that relevance and indicativeness agree around 73% of the time. This value can be considered as an assessor error rate to adjust the expected performance of sentence ranking methods using the TREC Novelty track assessments. Second, during our evaluations we discovered a bias that longer sentences were more likely to be judged as relevant. Our analysis demonstrates that this length bias in sentence ranking methods can lead to incorrect conclusions about the relative performance of sentence ranking techniques using query expansion. We then propose a novel evaluation of sentence ranking methods, which aims to isolate the sentence length bias. Using our enhanced evaluation method, we find that query expansion can effectively assist in the selection of short sentences.

We conclude our investigation with a second study to examine the effectiveness of query expansion approaches in query-biased summarisation methods to end users; that is, where top-ranked sentences are presented to users. While our previous evaluation suggests that query expansion approaches are beneficial for short sentences, our results indicate that subjects significantly tend to prefer query-biased summaries aided through expansion techniques approximately 60% of the time, for query-biased summaries comprised of short and middle length sentences. We suggest that our findings can inform the generation and display of query-biased summaries of IR systems such as search engines.

Chapter 1

Introduction

Many decisions in our daily life are based on information presented in the form of summaries. For example, reading the synopsis of a movie that can catch our interest, reading statistics about the benefits of sunscreen to protect the skin, or listening to traffic reports to avoid a crowded street, to mention a few. Summaries are succinct descriptions that emphasise key content from a source of information. In recent times, with the rise of search engines, one of the most widely used summaries is the "*snippet*". Snippets are short fragments of text extracted from a document, where query terms submitted by users to a search engine usually appear in these fragments.

Summaries have evolved mainly due to the large current volumes of information, diverse types of sources, the capabilities of retrieval systems to display and return information, and the seeking behaviours of users. Before the digital era, librarians assembled bibliographic catalogs of documents by providing descriptors that could help to characterise a document without reading it. These descriptors included the title of documents, key terms, an abstract, dates of publication, areas of interest, authors and many other fields found in bibliographic catalogs. Abstracts and titles were regarded as the most useful descriptors of documents [Barry, 1998; Janes, 1991; Marcus et al., 1978]. Since these abstracts were created by humans, the information flowed smoothly and coherently, capturing the general content of a document. However, the presence of an abstract could depend on its availability in a document, such as in research articles. In its absence, the leading sentences of a document were taken as extracts.

The automatic creation of extracts proliferated as emerging retrieval systems needed to manage larger volumes of information of a different nature. In addition, these retrieval

systems gave users autonomy to conduct their searches, so users could formulate requests without knowledge of how the information was organised. As users gained experience, they also became more demanding of the extracts returned. These extracts were therefore enhanced to contain information that had a relation with the submitted request, by presenting fragments of a text that matched query terms [Egan et al., 1989; Pedersen et al., 1991; Tombros and Sanderson, 1998]. Formally, these extracts are called query-biased summaries, which provide users with specific information tailored to their requests.

Modern retrieval systems, such as search engines, typically describe a document by presenting three main components: its title, a short query-biased summary (snippet), and a URL. According to the type of user requests, these three components can guide users to more easily identify helpful documents. There are different types of searches conducted on the Web. In a broad classification, requests can be: transactional, as users look on the Web to ask for services or resources; navigational, where users target for a particular Web site; and informational, where users try to locate specific information about a topic that can usually be satisfied by inspecting multiple documents [Broder, 2002].

The title, the snippet and the URL are usually encapsulated to form a textual caption [Clarke et al., 2007], as shown in Figure 1.1. Captions can be visually enhanced through images or small snapshots of the source. These extra elements can save users from having to read through textual captions. However, research indicates that textual features in captions are effective for users when looking for information that they have never seen before [Teevan et al., 2009] or related to an informational request [Al-Maqbali et al., 2010]. In particular, query-biased summaries can assist users more accurately for requests that need specific information. This is supported from studies conducted on archives [Fachry et al., 2010], newswire collections [Tombros and Sanderson, 1998], Web content [White et al., 2003], and evidence gathered from eye tracking techniques [Turpin et al., 2006].

This thesis studies techniques for the generation of effective query-biased summaries, specifically, the process of selecting informative sentences prior to displaying these as part of a query-biased summary. Sentences are useful information blocks to assemble summaries [Edmundson, 1969; Goldstein et al., 1999; Hovy and Lin, 1998; Jing et al., 1998; Kupiec et al., 1995; Luhn, 1958; Radev et al., 2004], since they convey single ideas and are usually suitable for space-limited presentation layouts to display summaries. We aim to construct query-biased summaries in the context of informational requests. That is, cases where a query-biased summary does not exactly contain a unique and definite answer to users requests as in Question Answering approaches. Rather, the summary points to documents of possible

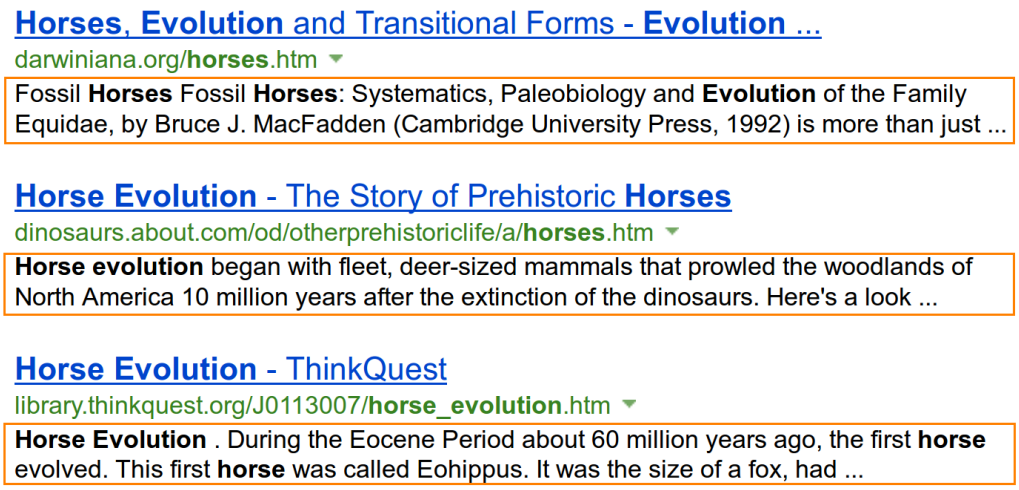

Figure 1.1: Example of captions displayed by a commercial search engine. The snippet component is enclosed in a box.

relevance to users.

1.1 Research Questions and Contributions

We investigate the following research questions.

- *How are query-biased summaries created by humans? How does this compare with current automatic summarisation methods?*

We aim to understand how humans construct different types of query-biased summaries. In particular, we study people while writing summaries (a *generative* approach) and selecting content from documents (an *extractive* approach) according to a specific information request. We use eye tracking techniques to examine strategies that people follow to construct their summaries. Eye movement data can help to determine the regions of a document that are read by people while creating query-biased summaries. We observe that previous research in automatic summarisation has not included eye tracking techniques, has generally focused on generic summarisation, or has explored human extractive approaches for query-biased summaries. However, limited effort has been conducted to explore both generative and extractive query-biased summarisation as it is carried out by humans.

We compare human query-biased summaries with automatic summarisation methods that rely on sentence ranking approaches. Based on eye tracking data, we aim to identify

parts of documents that people select (through generative or extractive summaries) that are ignored by automatic methods. We also propose to assess automatic methods based on a bag-of-words approach; that is, through vocabulary overlap between human and automatic summaries. We describe our user study and findings of human query-biased summaries in Chapter 3.

- *How to create effective query-biased summaries?*

From the eye tracking experiments above, we found a short query did not fully capture the subjects' internal model used to generate a query-biased summary. This issue is well known in retrieval as the "vocabulary mismatch" problem, where the actual words entered to initiate a search are concise, but may not be the same words used in a target document to describe the same information. Retrieval systems have employed query expansion techniques to overcome the lack of verbosity of users to define their requests, or prior knowledge about the information they are searching for. Query expansion aims to enrich an original request by automatically introducing terms that share a certain relation with the request such as frequency, co-occurrence or synonyms, for example. Previous research has investigated the effects of query expansion for summarisation purposes, and for passage retrieval. However, we observe that these studies do not explore query expansion with regards to query-biased summaries. For example, these studies do not focus on single document query-biased summarisation [Losada, 2010], employ detailed information requests [Sanderson, 1998], or simulate users' requests using the title of documents rather than employing formal testbeds [Amini et al., 2005; Han et al., 2000].

We propose to investigate the effectiveness of statistical query expansion in sentence ranking methods for query-biased summarisation. Specifically, we study the approach of Rocchio [1971] and Local Context Analysis [Xu and Croft, 1996; 2000], used for improving document retrieval. Expansion approaches explored in this thesis rely on sourcing additional terms from top ranked documents and from top ranked sentences. In addition to query expansion, we examine other document features employed in generic summarisation to improve the selection of sentences to generate query-biased summaries, namely clusters of significant words and sentence position.

Document ranking approaches are generally evaluated using the Cranfield methodology [Cleverdon, 1967], which employs a set of documents, information requests and relevance judgements. In order to assess the effectiveness of sentence ranking approaches, we employ sentence relevance assessments available in the TREC Novelty track by adapting the Cran-

field methodology to a sentence-based context. In Chapter 4, we describe the effectiveness of different sentence ranking methods, and the effectiveness of query expansion techniques applied to them.

- *How should one evaluate sentence ranking methods using sentence-level relevance data?*

The evaluation of summaries can involve intrinsic and extrinsic methodologies. In intrinsic approaches, automatic summaries are compared against model summaries created by humans. Consequently, the Cranfield-based evaluation resembles an intrinsic approach, since the set of relevant sentences can be considered a human summary. We propose to assess sentence ranking methods by inspecting the effects of sentence indicativeness and sentence length.

In Chapter 5, we hypothesise that not all sentences judged as relevant by TREC Novelty track assessors are good candidates to assemble a query-biased summary. That is, while they are topically relevant, this does not necessarily imply that they are indicative of the document content. We conduct a crowdsourcing experiment to explore whether sentences judged as relevant are also indicative, and through stochastic simulations we estimate how this affects the effectiveness of sentence ranking methods.

We detect that long sentences tend to be selected as relevant by TREC Novelty track assessors. Length bias is a problem that has been studied in document retrieval [Singhal et al., 1996; Losada et al., 2008], bibliographic catalog fields [Janes, 1991; Marcus et al., 1978] and passage retrieval [Callan, 1994]. However, it has not been explored for sentences in the context of sentence ranking for a query-biased summary task. We examine the relationship between the sentence length bias and sentence ranking methods assisted by query expansion, and propose a novel sentence ranking evaluation approach that isolates the sentence length bias in the collection assessments. These results are also examined in Chapter 5.

In extrinsic approaches, summaries are under the direct scrutiny of users performing tasks in a simulated work scenario. Our previous evaluations aim to gauge the effectiveness according to individual sentences as blocks of information. However, these sentences can be joined and presented to users as a final query-biased summary. In Chapter 6, we describe a crowdsourcing study to investigate whether participants prefer sentence ranking methods assisted by query expansion. That is, the top ranked sentences are compiled into a query-biased summary, which aims to assist users in identifying likely relevant documents. In particular, we compare sentence ranking methods assisted by query expansion with non-expansion

approaches. Further, we revisit the sentence length bias in query-biased summaries, investigating whether a summary constructed using long sentences is an effective substitute for more complicated query expansion techniques to minimise the vocabulary gap problem.

This thesis is structured as follows:

Chapter 2 reviews approaches for the generation of generic and query-biased summaries. This chapter surveys statistical query expansion approaches to obtaining extra terms such as relevance feedback and pseudo-relevance feedback techniques. In particular, we examine their use for summarisation purposes. We review methodologies to evaluate summaries, describe the testbed employed in our main experiments, and detail experimental approaches to collect user data such as eye tracking and crowdsourcing techniques.

Chapter 3 describes a user study that examines people while creating generative and extractive query-biased summaries. We explain our experimental design, and evaluate the performance of automatic sentence ranking methods based on eye tracking evidence and vocabulary overlap.

Chapter 4 explores effective sources of evidence to rank sentences and query expansion techniques, using sentence relevance assessments from the TREC Novelty track. The chapter investigates query expansion approaches that rely on top ranked documents and top ranked sentences.

Chapter 5 studies how to evaluate sentence ranking methods considering properties such as sentence indicativeness and sentence length bias in the TREC Novelty track data.

Chapter 6 describes the experimental setting of a crowdsourcing study to assess a set of ranked sentences as the final query-biased summary of a given document. We investigate whether participants find query-biased summaries constructed using query expansion techniques more useful.

Chapter 7 summarises the contributions and findings of this thesis, and provides recommendations for future work.

Chapter 2

Background

Prior to examining the state-of-the-art in automatic text summarisation, we provide a review of typical pre-processing document approaches, and retrieval models in the first section of this chapter. We assume that an IR system returns documents, which will then be summarised.

We describe generic extractive summarisation methods, and then we survey practices for the creation of query-biased summaries. In particular, we investigate sentence ranking approaches as an initial step to assemble query-biased summaries, and query expansion techniques to improve the selection of indicative sentences from a document. The chapter also discusses evaluation approaches to measure the effectiveness of summarisation methods, the testbed employed in our experiments, as well as eye tracking and crowdsourcing techniques.

2.1 Before Summarisation

Information retrieval systems serve users to locate documents that have a certain similarity with their information needs. Before returning an answer, IR systems generally pre-process documents by identifying tokens, case-folding terms, removing stopwords, or stemming terms. These practices enable the collection of vocabulary statistics before employing a retrieval model to score documents. These documents are ranked by first showing those that have a higher similarity with user requests. In this section, we explain typical practices to pre-process documents, and two widely employed retrieval models.

The construction of summaries can be carried out prior to query time, so summaries are stored with a reference to each document. These static summaries can cover general aspects of information in that document. Another approach is to dynamically assemble summaries at query time; thus, they are focused on requests submitted by users. This is a common

scenario for retrieval systems such as modern search engines. This thesis does not investigate document retrieval aspects; that is, we conduct a summarisation process independent from retrieval. However, we use pre-processing approaches and retrieval models in experiments conducted in Chapter 4.

2.1.1 Pre-Processing of Documents

The pre-processing of documents can involve the following treatments: tokenisation, case-folding, stemming, and stopping. Depending on the application or language, some of these treatments may not be carried out.

Tokenisation. Tokenisation is the process of identifying elementary units of content within a text. These units, called *tokens*, can be represented as single words delimited by the space character. This assumption can effectively work for some languages such as English, Spanish, and Italian. For Asian languages the tokenisation process is not as trivial as other languages; however, this is out of the scope of this thesis.

Stemming. Words are constructed with different elements to represent a new term meaning or grammatical function [Yule, 2010]. These elements are known as *morphemes*, and can be classified as free or bound morphemes. Free morphemes, also called *stems*, represent a concept and are regularly used to name nouns, adjectives and verbs. Bound morphemes are shorter and specify the functionality of a free morpheme, such as plural, tense, or noun. For example, the terms `engines`, `engineering` and `engineer` are formed by the free morpheme *engine*, and three bound morphemes *-s*, *-ing* and *-er*. Bound morphemes can be linked at the beginning or at the end of a stem in the form of prefixes or suffixes, respectively. For instance, the term `undetected` has the prefix *un-* and the suffix *-ed*, while the stem is *detect*.

Stemming is the process of removing suffixes, since it can be straightforwardly automated through a set of rules based on pattern suffixes to reduce words to their stems [Porter, 1980]. Stemming typically does not remove prefixes as these are more complex to detect. Words such as *unary* or *underground* lexically match the prefix *un-* at the start of the words; however, the meaning of these terms does not convey the lack or negation of something, which is associated to the particle *un-*. The stemming process is crucial when creating the vocabulary of a given collection, as it reduces the number of entries in a lexicon. From the example above, the stem *engine* records a frequency of three, instead of three different words with frequency of one. Due to the diversity of language, pattern-rule stemmers can reduce

stems incorrectly. Other stemming algorithms which rely on linguistic features of words have emerged to alleviate these shortcomings [Krovetz, 1993].

Stopping. Free morphemes are classified as lexical and functional [Yule, 2010]. Lexical morphemes include nouns, adjectives and verbs, which describe the content of a document. Functional morphemes provide the cohesive structure of discourse; these can be found as conjunctions, prepositions, articles and pronouns. Functional morphemes and highly frequent words in a collection compile *stopword* terms. Stopping is the process of ignoring stopwords, since they may not help to describe the content in a document. However, stopping may be harmful while processing some queries or collections. For example, consider a user looking for information about the Beatles' song: "here, there and everywhere". A stopping process may produce an empty string as query, if these terms were in a stopword list. Thus, some applications such as Web search engines preserve stopwords at indexing time [Baeza-Yates and Ribeiro-Neto, 1999]. In this thesis, we do not deal with this type of queries, and stopwords were removed from requests and documents while ranking sentences, as described in Sections 3.4.2 and 4.2.

2.1.2 Retrieval Models

Retrieval models aim to measure the similarity between user requests and documents in large collections. The Boolean model, one of the earliest approaches, relies on Boolean operators to represent queries to a system, and set theory to locate documents containing query terms [Baeza-Yates and Ribeiro-Neto, 1999]. However, users can experience difficulties to formulate complex requests through Boolean operators. Moreover, the occurrence of query terms in a document is represented on a binary basis regardless of the term frequency distribution in the document. These shortcomings led researches to investigate more sophisticated similarity functions that rank documents according to term frequency or probabilistic distributions. Particularly, we describe the Vector Space Model and the Okapi BM25 ranking functions, as these were employed in this thesis. However, a variety of retrieval models have been proposed in the literature, see for example Baeza-Yates and Ribeiro-Neto [1999], Croft et al. [2009] or Büttcher et al. [2010] for a wider review.

The Vector Space Model

A method that measures document similarity based on a term frequency distribution is the Vector Space Model (VSM) [Salton et al., 1975]. The model represents documents and queries as weighted vectors. Therefore, the angle between both vectors can be seen as a measure that quantifies how similar documents are with respect to a query. The VSM uses the cosine (C) between vectors, supported by a term-weighting approach. The TF*IDF approach is widely employed to assist the computation of document-query similarities [Salton and Buckley, 1988; Zobel and Moffat, 1998]. The term frequency, TF, identifies terms that may determine the content of a single document. It is given as $f_{d,t}$ and corresponds to the number of occurrences of term t in document d. The TF component is defined as:

$$TF = 1 + \ln(f_{d,t}) \tag{2.1}$$

The inverse document frequency, IDF, is the reciprocal of the count for the number of documents in which term t appears (f_t) in a collection of N documents. This is based on the observation that common terms are not good to distinguish useful documents in a whole collection. The IDF is instantiated as:

$$IDF = \ln\left(1 + \frac{N}{f_t}\right) \tag{2.2}$$

Thus, the cosine between of a query vector Q and a document D assisted by the TF*IDF term-weighting approach is defined as:

$$C(D,Q) = \frac{\sum_{t \in Q \cap D}(1 + \ln(f_{d,t})) \ln\left(1 + \frac{N}{f_t}\right)}{\sqrt{\sum_{t \in D}(1 + \ln(f_{d,t}))^2}} \tag{2.3}$$

The VSM has been widely employed given its simplicity to compute, and its robustness as a baseline [Baeza-Yates and Ribeiro-Neto, 1999]. A limitation of the model is term-independence; that is, no term dependency is taken into account. For example, the query "*Olympic Games*" is broken in two tokens for searching for documents containing such terms. The VSM model then can return documents related to other types of games such as "video games" or "hunger games", or documents that separately mention the terms "games" and "olympic".

The Okapi BM25 Ranking Function

The Okapi BM25 ranking function is the product of a series of theoretical approaches that rely on estimating the probability that a document is relevant to a query [Robertson et al., 1995; Spärck-Jones et al., 2000]. A probabilistic model can assume a relevance feedback approach (a user identifying relevant documents after submitting a query) to represent retrieval in a term-weighting schema similar to the TF*IDF. Robertson and Spärck-Jones [1976] proposed a probabilistic version of the IDF component as follows:

$$w_t = \log \frac{(r_t + 0.5)(N - f_t - R + r_t + 0.5)}{(f_t - r_t + 0.5)(R - r_t + 0.5)} \tag{2.4}$$

where R is the number of relevant documents given a query, and r_t is a subset of relevant documents that contain the term t. Since relevant documents are unknown at the moment of issuing a query, initially $R = 0$ and $r_t = 0$. Thus, w_t can be simplified as:

$$w_t = \log \left(\frac{N - f_t + 0.5}{f_t + 0.5} \right) \tag{2.5}$$

The Okapi BM25 function incorporates a TF component, which depends on the document size. A parameter K in the function varies according to the document size (dl), and the average length of documents in the collection ($avdl$). The parameter K is defined as:

$$K = k_1 \cdot \left((1 - b) + b \cdot \frac{dl}{avdl} \right) \tag{2.6}$$

Thus, the Okapi BM25 function is defined as:

$$BM25(Q, D) = \sum_{t \in Q} w_t \frac{(k_1 + 1) f_{d,t}}{K + f_{d,t}} \cdot \frac{(k_3 + 1) f_{q,t}}{k_3 + f_{q,t}} \tag{2.7}$$

where the parameters k_1, k_3 and b have been established through exploring optimal values in TREC experiments [Croft et al., 2009]. In particular, often k_1=1.2 and b=0.75, while k_3=0 as it can be assumed that query terms are not repeated in a request such as Web queries, which tend to be short.

In the next sections, we provide a review of automatic summarisation, extractive summarisation, query-biased summarisation, query expansion approaches, summary evaluation, evaluation of sentence ranking methods, and techniques to collect user data.

2.2 Automatic Summarisation

Summaries are abbreviated versions of documents that aim to highlight important information. There are multiple factors that shape the focus and content of a summary such as functionality, construction approach, audience and domain, to mention a few [Spärck-Jones, 2007]. Based on functionality, summaries can be *informative* or *indicative* [Edmundson, 1969]. Informative summaries, also known as generic summaries, provide a global idea about the document content and are usually comprised of paragraphs [Salton et al., 1997; Strzalkowski et al., 1998] or several sentences [Brandow et al., 1995; Goldstein et al., 1999; Hovy and Lin, 1998; Jing et al., 1998; Kupiec et al., 1995; Luhn, 1958]. In contrast, indicative summaries serve to identify useful documents among large text collections. In some cases the length of indicative summaries does not exceed 2 sentences or 200 characters [Buchheit, 2005; Gomes and Smith, 2003], and their construction can be biased towards requests of users [Tombros and Sanderson, 1998; White et al., 2003].

In terms of the construction approach, we identify *abstracts* and *extracts*. Humans are able to detect representative content in a document by creating new prose, and by merging two single ideas into one more complex while summarising a text [Brown and Day, 1983; Irwin and Doyle, 1992; Winograd, 1984]. Abstractive summarisation attempts to automatically achieve these cognitive processes by relying on techniques used in Artificial Intelligence [Fum et al., 1982] or Natural Language Processing [Barzilay and Elhadad, 1997; Marcu, 1997]. On the other hand, extractive summarisation does not generate new text; rather it takes verbatim parts from source documents that are deemed to be important for a summary. Constructing extracts consists of scoring passages (paragraphs or sentences) within a document, and assembling those highly ranked passages to make up a summary.

Another two factors that affect the creation of summaries are the audience and the domain. Specific audiences may require more elaborate summaries that help users (with a defined profile, expertise, or specialization level) to complete particular tasks [Afantenos et al., 2005; Fiszman et al., 2004]. For example, a biologist who searches in a medical database for a particular gene would require summaries that reveal the chemical composition, mutations or associated diseases to that gene. In this case, the domain is related not only to the area of application, but also to a specialised audience. However, summaries can be created for general audiences, where users do not have a defined profile, or may lack familiarity with a topic.

Given these four factors, functionality, construction approach, audience and domain, this thesis focuses on extractive approaches for creating indicative summaries directed to a general audience and domain. Specifically, we study the generation of short summaries that are tailored according to query terms provided by users, also known as *"query-biased"* summaries, as a sentence ranking problem. In Section 2.3.1, we introduce typical approaches for creating generic summaries. While these techniques do not involve users requests, they can support the assembly of query-biased summaries. Query-biased summarisation, which is under study in this work, is detailed in Section 2.4. For an extensive review on other summarisation approaches see for example Lloret and Palomar [2012]; Nenkova and McKeown [2011]; Paice [1990]; Saggion [2008] and Spärck-Jones [2007].

2.3 Extractive Summarisation

Extractive summarisation scores passages of documents based on certain patterns or attributes. These patterns or attributes guide the identification of salient content in documents. Generally, extractive approaches select the top m passages and present them as a summary. Previous research has identified three types of passages: semantic, window and discourse [Callan, 1994]. However, not all of them are appropriate for the generation of summaries.

Semantic passages represent areas in a document that cover the same topic. Nevertheless, documents that are short and lack structure may lead to inaccurate detection of topical areas [Hearst, 1997]. Window passages are constructed by a stream of text at a specific length cut off, which can be given in words or characters. Discourse passages can be distinguished by specific tokens in the text such as punctuation marks or the new line character. Thus, the identification of window or discourse passages in a document is relatively straightforward. In the context of extractive summarisation, research has focused on discourse passages, since these are simple to determine and lead to coherent pieces of text instead of incomplete or fragmented information, as window passages can provide.

Within a document, discourse passages take the form of sections, paragraphs and sentences, with the latter two being better suited for summarisation tasks. Paragraphs convey more information and have been applied for generic summaries [Salton et al., 1997; Strzalkowski et al., 1998]. Sentences are typically preferred for assembling short summaries due to their capability to present single and complete ideas. In this thesis, we employ sentences

to create query-biased summaries as shown in Chapter 6. Sections 2.3.1 and 2.4 examine extractive approaches to create generic and query-biased summaries, respectively.

2.3.1 Generic Summarisation

Generic summarisation can use information directly collected from documents or from shallow sentence attributes to rank sentences. Information that can be gathered from documents typically includes: determining significant terms from word frequency statistics; identifying words of titles and headings; or using metadata information and anchor text. In contrast, shallow sentence attributes are independent from the document vocabulary, and employ features such as the ordinal position of sentences and sentence length. Summarisation methods can merge information gathered from documents and shallow sentence attributes to improve effectiveness. In particular, these approaches are used in linear combination, where constants tune the value that each feature contributes to rank a sentence [Radev et al., 2004; Tombros and Sanderson, 1998; Turpin et al., 2007; White et al., 2003]. A linear combination approach is investigated in Section 4.4.1. In following sections, we describe approaches based on document content and on shallow sentence attributes.

Significant Words

The significance of words can be attributed to the frequency with which they appear within a document, since these terms may help to define the content of documents [Edmundson and Wyllys, 1961; Edmundson, 1969; Luhn, 1958]. Luhn [1958] hypothesised that highly frequent terms (generally stopwords) and infrequent terms can be labelled as non-significant words, while the remaining terms with a mid-frequency distribution of occurrence compile a set of significant words. Thus, sentences are comprised of clusters of significant and non-significant terms [Luhn, 1958]. Clusters in sentences start and end with a significant word; however, gaps between them of at most four non-significant terms are allowed [Edmundson and Wyllys, 1961; Luhn, 1958]. Given that a sentence may contain several clusters, the highest cluster weight determines the sentence score. A cluster sentence score is defined as:

$$CL_s = \text{argmax}_{c_j} \left(\frac{|\text{significant terms in } c_j|^2}{|c_j|} \right) \tag{2.8}$$

where c_j is cluster j in sentence s, and the notation $|x|$ represents the number of terms in that cluster.

Luhn relied on choosing words that were within specific frequency thresholds. The setting of such thresholds may require the analysis of multiple text collections to obtain optimal values. Tombros and Sanderson [1998] studied a small random sample of documents from the Wall Street Journal (1998) in order to establish a simple formulation to discriminate non-significant terms. Their empirical heuristic depends on the number of sentences in a document, s_d. Specifically, a term t is taken to be significant if its frequency in a document, $f_{d,t}$, exceeds or equals one of the following restrictions:

$$f_{d,t} \geq \begin{cases} 7 - 0.1 * (25 - s_d) & \text{if } s_d < 25 \quad (1) \\ 7 & \text{if } 25 \leq s_d \leq 40 \quad (2) \\ 7 + 0.1 * (s_d - 40) & \text{if } s_d > 40. \quad (3) \end{cases} \tag{2.9}$$

For example, for a document containing 18 sentences, this approach considers significant terms to be those occurring at least six times, as this satisfies the first restriction in Equation 2.9. Given the simplicity with which significant terms can be identified, this approach has been employed to construct Web query-biased summaries [Wang et al., 2007] and efficient methods for snippet generation [Turpin et al., 2007]. We examine the effectiveness of significant term heuristics further in Chapter 4.

A more formal term-weighting approach to gather significant terms uses techniques from document retrieval such as the TF*IDF approach [Salton and Buckley, 1988] (see Section 2.4.1). TF*IDF values are computed for a target document or clusters of documents (in the case of multi-document summaries [Radev et al., 2004]), and the top n scoring terms or terms with a weight above a specific threshold are assumed to be significant [Brandow et al., 1995; Hovy and Lin, 1998; Teufel and Moens, 1997].

Cue Words and Cue Phrases

Cue words and cue phrases identify sentences that may contain relevant information within a paragraph or within a section of a document [Edmundson, 1969; Teufel and Moens, 1997]. Cue words can be compiled in dictionaries and glossaries. Cue words in dictionaries are gathered from collection statistics, while glossaries are obtained from a document content. Cue words receive positive or negative weights depending on the meaning they convey [Edmundson, 1969; Teufel and Moens, 1997]. For example, words such as *"crucial"* and *"unsuccessfully"* can be assigned with positive and negative weights, respectively. Therefore, sentences containing such terms denote information of special consideration to analyse, or content that

can be ignored instead. The main drawback of cue words is the weighting process, as positive or negative values of words need to be given by subjects, which requires intensive human intervention.

On the other hand, cue phrases are short sentences or noun-phrases that provide a preamble to specific sections in a document. These phrases emphasise the aim of an information block, or indicate the transition to a new topic in a document. For instance, common cue phrases include *"This paper aims to", "To sum up",* or *"In conclusion"*. While cue phrases can be easily identified in documents with a well-defined structure, their gathering may pose a challenge when documents lack organisation and signalling vocabulary. Given the shortcoming of cue words and cue phrase approaches for scoring sentences, these are not studied in this thesis.

Title and Heading Words

Title and heading words are useful to detect topical terms that may describe the content of a document and specific sections, respectively. These terms are provided by the author of a document, who potentially is the best person to supply key terms through title and headings. Therefore, summarisation methods can rank sentences according to the occurrence of these terms [Edmundson, 1969; Joho et al., 2008; Tombros and Sanderson, 1998; White et al., 2003]. We do not investigate this approach, since title and heading terms may not have any relation with the document at all or with the specified information request.

Metadata and Anchor Text

Other sources of information available in documents that have been used for summarisation purposes are metadata information and anchor text, particularly in text collections containing Web content. Search engines have taken advantage of metadata information (if available) for displaying snippets [Kaisser et al., 2008]. Metadata is encapsulated in the HTML document through markup language;[1] however, it is not displayed in the browser, so is not visible to the user. Usually, metadata information is not automatically generated; rather it is provided by humans. For instance, we observe that after submitting the query "*health insurance*" to a commercial search engine the snippets in Figure 2.1 were assembled by displaying the information in the tag: `<meta name="description" content= />`. While this approach does not involve a sentence ranking process to construct snippets, it overcomes the lack of

[1] `http://searchenginewatch.com/article/2067564/How-To-Use-HTML-Meta-Tags`

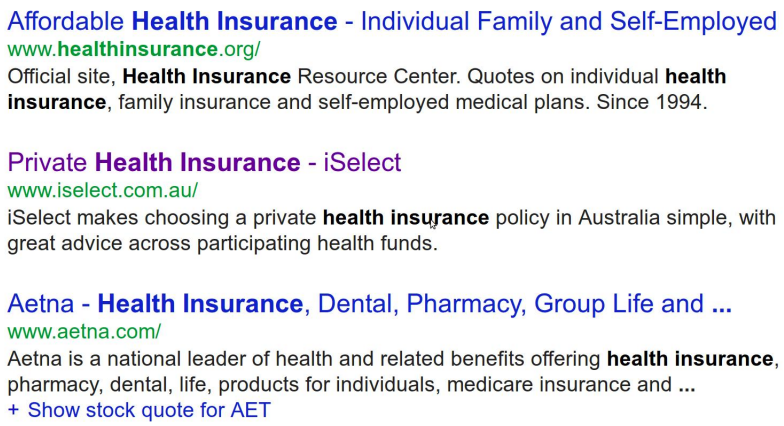

Figure 2.1: Example of metadata information used for displaying Web search snippets.

informational content in some Web documents given the amount of scripting language or advertising material.

Anchor text is another element available in Web content employed for assisting Web site retrieval [Craswell et al., 2001]; however, it can also be used for summarisation tasks. The anchor text is a brief description of a link that points to another Web document (or resource). This description is provided by a person who has an understanding of the target document. Terms in anchor text can be grouped to construct a summary with the appearance of an abstract [Amitay and Paris, 2000], or be used to score sentences [Wang et al., 2007].

In this thesis, we do not investigate metadata or anchor text components to create query-biased summaries, since it is outside the scope of this thesis. For example, metadata is likely to be more helpful to construct summaries for navigational requests rather than informational requests. In addition, anchor text is not always descriptive; it is common to find anchor text such as "`Click here`", "`More info`", or "`Find it here`".

Sentence Position

The position of a sentence in a text, a feature independent from document content, may indicate good candidate sentences for summaries. Edmundson [1969] investigated whether the relevance of a sentence can be estimated by its location in a document. For example, newspaper articles briefly answer the questions *what, who, where* and *when* in the first sentences of an article. Radev et al. [2004] used the inverse position that sentences have in a document to complement sentence scores for a multi-document summarisation approach. That is, the importance of a sentence decreased according to its location in a document [Ko

et al., 2008; Radev et al., 2004].

In particular, leading sentences of a document have been shown to be more useful than sentences further in a text, as these can provide an overall view of the content [Brandow et al., 1995; Edmundson, 1969; Goldstein et al., 1999]. This approach was convenient in early search engines such as AltaVista [Berger and Mittal, 2000], since leading sentences could be easily stored at indexing time and subsequently displayed as summaries of documents. Turpin et al. [2007] proposed a position-biased score for pruning documents as follows:

$$POS_s = \begin{cases} 2, & \text{if } s \text{ is the first sentence in a document} \\ 1, & \text{if } s \text{ is the second sentence, or the title} \\ 0, & \text{otherwise.} \end{cases} \tag{2.10}$$

We investigate the effectiveness of this approach for the construction of query-biased summaries in Section 4.4.1.

Sentences can also be examined relative to their position in a paragraph [Baxendale, 1958; Lin and Hovy, 1997]. By employing the Ziff-Davis newspaper collection, Lin and Hovy [1997] found that the first sentences in a paragraph often contain key terms. These sentences are likely to be more informative, as they define the thesis or topic in a paragraph. However, they discovered that this assumption was not applicable for ending sentences in a paragraph. Ending sentences can provide salient content because they conclude a block of information, which depends on the writing style of authors. Other studies have demonstrated that, due to multiple writing styles, topic sentences can occur in any part of a paragraph or simply not appear [Braddock, 1974; Donlan, 1980]. We note that work conducted to identify topic sentences is scarce [Baxendale, 1958; Braddock, 1974; Donlan, 1980], and more research is required to identify topic sentences in modern text collections.

In Web content, paragraphs cannot be explicitly defined due to the flexibility of markup language to not include the tag `</p>`, which closes a block of information. In Section 4.2.3, we study the position of a sentence in regards to its location within a document as a mechanism to resolve tie scores between sentences, since the identification of paragraphs is not always available in text collections.

Sentence Length

Long sentences are more likely to contain significant terms, cue terms, cue phrases, and title terms compared to short sentences. This increases the probability of long sentences being

selected and included in a summary. In order to avoid a bias factor introduced by length, sentence scores can be normalised by dividing them by the total number of words in each sentence [Tsegay et al., 2009]. Another approach is to ignore brief sentences, since they might not include relevant content [Kupiec et al., 1995]. We study the length feature as a complement in sentence ranking methods in Chapter 5, and its effects on the effectiveness of these methods.

Long sentences are prone to include words that do not contribute to the relevance of a sentence. Natural Language Processing approaches indicate that sentences have fragments that provide crucial information (*nuclei*), while others serve as complementary information (*satellites*) [Grefenstette, 1998; Jing and McKeown, 1999; Knight and Marcu, 2002; Sporleder and Lapata, 2005]. Thus, removing satellites such as non-verb clauses may not harm the central idea, and result in a compressed sentence. However, the removal of satellites can lead to poor coherence or readability given the lack of context for sentences. For these reasons, we do not explore sentence compression approaches. For instance, consider the following sentence, segmented in three clauses, related with the extinction of pandas. Clause 2 is a non-verb clause that provides certain context about the role of *Pan Wenshi*, and specifies that pandas are in danger of starving.

Clause 1: (nuclei)	Clause 2: (satellite)	Clause 3: (nuclei)
Pan Wenshi of Beijing University,	*one of the world's foremost authorities on pandas,*	*said the animals were never in danger of starving.*

2.4 Query-Biased Summarisation

Query-biased summarisation — also called *query-dependent, query-specific* or *query-relevant* — is a type of extractive summarisation that favours the selection of passages containing query terms. In large text collections, query-biased summaries help users to ignore irrelevant documents, and to inspect those that appear to be relevant [Tombros and Sanderson, 1998; White et al., 2003]. For example, commercial search engines display *snippets* of returned documents, where a specific character-based window surrounding query terms can be shown as summaries [Gomes and Smith, 2003]. Users of retrieval systems can employ query-biased summaries to guide their searches without inspecting whole documents. Thus, the functionality of query-biased summaries is to indicate relevant documents [Edmundson, 1969].

We argue that query-biased summaries are different from question answering approaches or personalised summaries. Question answering aims to provide a definite response to a given request as single facts within the content of a summary [Monz, 2003; Murdock, 2006; Voorhees, 1999]. For instance, *"Who was the richest man in 2012?"* or *"How many calories are there in a Big Mac?"* are natural language questions with a unique response that question answering can address. In personalised summaries, users fill templates to specify their preferences or characteristics of the summary [Fum et al., 1982; Radev and McKeown, 1998]. Users can indicate the desired summary length, or the topics they are interested in seeing as the synopsis of a document [Berkovsky et al., 2008; Díaz and Gervás, 2007].

We observe that question answering and personalised summarisation methods employ more elaborate and eloquent inputs to create summaries. This thesis studies query-biased summaries as a sentence ranking approach, where concise descriptions of requests set a challenge to construct an indicative summary of a document. We focused on informational requests, which require users to find specific information that can be spread through multiple documents to satisfy their request [Broder, 2002]. Query-biased summarisation methods investigated in this thesis score sentences relying on document ranking functions, and simple heuristics that count query term occurrences. Both approaches are examined in the next two sections. This thesis studies statistical approaches for the construction of query-biased summaries; however, we briefly survey machine learning, clustering and language models in Section 2.5 as alternative techniques to generate summaries.

2.4.1 Document Ranking Functions

A variety of document retrieval models have been proposed in the literature such as the Vector Space Model [Salton et al., 1975], the Okapi BM25 ranking function [Robertson and Spärck-Jones, 1976; Robertson et al., 1995] and Language Models [Croft et al., 2009]. Therefore, by treating each sentence as a "document" it is straightforward to apply these ranking functions to score sentences relative to a query for sentence retrieval tasks [Allan et al., 2003; Losada, 2010], or summarisation [Goldstein et al., 1999; Han et al., 2000; Hovy and Lin, 1998; Varadarajan and Hristidis, 2006]. For example, the cosine similarity in the Vector Space Model calculates the Euclidean distance between weighted vectors of documents and a given query, as explained in Section 2.1.2. Thus, a short distance between both vectors indicates that a query shares a high similarity with a sentence. Allan et al. [2003] adapted

the VSM method for ranking sentences towards a query as:

$$R(s|q) = \sum_{t \in q} \log(f_{t,q} + 1) \log(f_{t,s} + 1) \log\left(\frac{n+1}{0.5 + sf_t}\right) \tag{2.11}$$

where $f_{t,q}$ and $f_{t,s}$ is the occurrence of term t in query q and sentence s, respectively. The number of sentences in the collection being scored is given by n, and sf_t is the number of sentences in which the term t appears.

Previous research has not reported significant differences when the VSM was compared with the Okapi BM25 ranking function [Allan et al., 2003; Losada, 2010] or Language Models with the Kullback-Leibler divergence [Allan et al., 2003; Li and Croft, 2005; Losada, 2010]. Thus, we study effectiveness of the VSM adaptation proposed by Allan et al. [2003] for ranking sentences in Chapters 4 and 5.

2.4.2 Query Term Occurrence Approaches

Simple and less formal heuristics to rank sentences depend on counting occurrences of query terms in such sentences. Tombros and Sanderson [1998] introduced a scoring approach that relies on the appearance of query terms, which is computed in a similar fashion as the clusters of significant words proposed by Luhn [1958]. The query-biased score for a sentence is defined as:

$$QB_s = \frac{(qt)^2}{q} \tag{2.12}$$

where qt is the number of unique query terms that occur in sentence s, and q is the total number of words in a query string. This heuristic has been applied for the construction of Web page summaries [Wang et al., 2007; White et al., 2003]. We study the effectiveness of this approach in Chapters 4 and 5.

Other query term occurrence approaches include the counting of repeated query terms and the longest contiguous sequence of query terms in a sentence to efficiently generate short summaries [Turpin et al., 2007]. In commercial applications, query term occurrence is used for extracting "query-relevant" parts from documents [Gomes and Smith, 2003]. These extracted parts are not only useful for displaying snippets of search engines, but also for detecting duplicate documents without analysing an entire document.

2.5 Other Summarisation Approaches

This section briefly reviews other approaches that construct generic and query-biased summaries; however, these were not investigated in our work.

2.5.1 Machine Learning

Machine learning can be defined as the collection of techniques that identify patterns or features from input training data. The "learned" patterns are then applied to test data. Machine learning has been employed to choose relevant sentences by using classifiers [Kupiec et al., 1995; Mani and Bloedorn, 1998; Teufel and Moens, 1997] or learning to rank methods [Amini et al., 2005; Metzler and Kanungo, 2008; Wang et al., 2007]. Kupiec et al. [1995] addressed the selection of sentences as a classification problem, that is, relevant sentences were separated from non-relevant. Using a collection of scientific articles, they showed that a Bayesian-based classifier improved the creation of generic summaries. Such a classifier employed shallow sentence features (length and position) and cue words. In contrast, learning to rank approaches have been studied for query-biased summarisation of Web content [Wang et al., 2007], scientific papers [Amini et al., 2005] and newswire collections [Metzler and Kanungo, 2008].

The effectiveness of machine learning approaches can be described as limited for the following reasons. Experiments conducted using the TREC Novelty track data (detailed in Section 2.9) showed that the selection of features to find relevant sentences were not robust among collections [Metzler and Kanungo, 2008]. In other cases, a simple linear combination of features was only slightly behind Support Vector Machines in terms of effectiveness [Wang et al., 2007]. Other work has used the title of documents [Amini et al., 2005] or the top-frequent terms in documents selected by users [Mani and Bloedorn, 1998] to form a set of candidate query terms. However, the effectiveness of these machine learning approaches has not been explored employing real user request or formal testbeds.

2.5.2 Clustering and Document Graphs

Clustering is a technique that aims to group elements either by their similarities or by their relations [Hartigan and Wong, 1979; El-Hamdouchi and Willet, 1989]. Elements in each cluster, in this case sentences or paragraphs, are assumed to cover the same topical content.

Therefore, a summary can be comprised of the most representative sentence in each cluster [Erkan and Radev, 2004; Hatzivassiloglou et al., 2001]. In sentence ranking approaches, the top n clusters [Kallurkar et al., 2003] can be used instead. However, these summarisation methods depend on an effective clustering technique capable of determining an optimal number of clusters.

Regarding document graphs, Salton et al. [1997] proposed to construct graph-like structures, called *relationship maps*, by computing similarities among paragraphs of encyclopedic articles. They found that the most connected paragraphs could be used to create generic summaries. However, as we have discussed in Section 2.3, paragraphs may not be suitable units of extraction for assembling short query-biased summaries. Another application of graph-like methods can be found in multi-document summarisation [Erkan and Radev, 2004]. Varadarajan and Hristidis [2006] employed a similar approach of graph-like structures as Salton et al. [1997] adapted for sentences called *document graphs*. From these document-graphs, spanning trees involving query terms were selected to form a query-biased summary. Possible shortcomings of this technique include the extra capacity required to store document graphs previous to the retrieval, and the cost of identifying optimal trees at query time.

2.5.3 Language Models

Berger and Mittal [2000] proposed that the construction of query-relevant summaries not only depends on user requests, but also on information that depicts the general content of a document. They proposed to use a unigram language model to study the probability distribution of terms and query terms in a document. In order to evaluate this approach, they employed "Frequently Asked Questions", where a question may be similar to an information request and its answer would represent an ideal summary.

Other techniques that may assist to construct query-biased summaries are based on translation models [Murdock and Croft, 2005] and relevance models [Balasubramanian et al., 2007]. The former has been shown to be effective for question answering tasks, and the latter has improved the detection of redundant content. We do not detail such approaches as question answering and novelty detection are outside the scope of this thesis.

2.5.4 Word Formatting

Word formatting such as bold, italics, underline and color can be an indication of the emphasis that a term conveys. Authors employ word formatting to capture the attention of the reader in specific areas of text. Word formatting relies on markup language to delimit the areas of text that receive one or more formatting styles. Web query-biased summarisation methods have applied a weight to sentences containing formatted words [Verstak and Acharya, 2012; White et al., 2003]. However, the style formatting of a word is subject to the author and there is no rule that indicates the importance level of each format. Thus, we did not explore this approach in our work.

2.6 Query Expansion

Users describe their information requests to IR systems by reducing them to key terms, which ideally help to retrieve a useful set of results. The typical interface for initiating a search consists of a text area, where users are typically far from eloquent in depicting details of the information they are seeking. Previous research has found that descriptions provided by users are short [Bendersky and Croft, 2009; Jansen et al., 2007] and inaccurate [Furmas et al., 1987], these factors reduce the probability of an IR system being able to locate helpful results [Buckley, 2004]. This is the typical vocabulary mismatch problem (or gap) between users and content documents. Users may lack previous knowledge or are unfamiliar about a topic. These factors lead users to select inadequate words to define their information needs [Rocchio, 1971]. On the other hand, authors of documents may write without a specific audience in mind, use synonyms to avoid repetitive content, or employ a specialised lexicon. One approach to reducing the vocabulary gap and to boost the effectiveness of IR systems is by using query expansion. Query expansion techniques automatically introduce additional terms to the original query, aiming to increase the set of relevant results. For example, the information request "*cloning of Dolly*" can be expanded with terms related to the topic such as "*laboratories, genetics, research*" or "*cells*".

Query expansion has been extensively studied for document retrieval [Billerbeck, 2005; Cao et al., 2008; Carpineto and Romano, 2012; Efthimiadis, 1996; Jing and Croft, 1994; Rocchio, 1971; Salton and Buckley, 1990; Xu and Croft, 1996]. In this thesis, however, we investigate its application to enhancing sentence ranking to construct query-biased summaries. As mentioned above, information supplied by users may be insufficient for returning relevant documents. For a summarisation approach a concise query reduces the capability of

automatic methods to assemble query-biased summaries that point to potentially relevant documents to users. For instance, a query term can occur several times in a document; however, it is less likely to appear in a sentence twice or more.

Query expansion techniques can include: *relevance-feedaback* and *pseudo-relevance feedback* to automatically enlarge a query. We briefly describe these approaches and outline their application to the summarisation problem, with specific emphasis on query-biased summaries. For a wider review of query expansion techniques in document retrieval see for example Billerbeck [2005], Carpineto and Romano [2012] or Efthimiadis [1996].

2.6.1 Relevance Feedback

During a searching session, a user submits a query to an IR system, and it returns a list of likely relevant documents. These documents are sorted according to their highest similarity to the request. Users then can "feed" the retrieval system by selecting documents from the list of results which presumably are relevant to the request. The chosen documents are then used by the system to gather extra terms, which will be used to expand the original query. This approach of query expansion is known as relevance feedback [Rocchio, 1971; Salton and Buckley, 1990].

Early relevance feedback techniques focused on generating an *"optimal query"*, which was intended to boost the selection of terms in relevant documents and to reduce the probability of weighting terms in documents of low significance. Rocchio [1971] modelled a relevance feedback technique to construct such an optimal query through a parameterised function that required identification of relevant and irrelevant documents as follows:

$$Q_{\text{optimal}} = \frac{1}{|R|} \sum_{d \in R} d - \frac{1}{N - |R|} \sum_{d \in (C-R)} d \tag{2.13}$$

To obtain an optimal query requires to firstly identify all relevant (R) and irrelevant documents ($N - R$) in a collection C of N elements: an impractical approach for large collections. As a solution to this inconvenience, after issuing an initial query Q_0, relevance feedback was collected by asking users to assess a subset of relevant documents (R') and non-relevant documents ($\bar{R}'$) from a ranked set of documents. A weighted query vector Q_1 was returned representing the enhanced query, defined as:

$$Q_1 = \alpha \times Q_0 + \frac{\beta}{|R'|} \sum_{d \in R'} d - \frac{\gamma}{|\overline{R'}|} \sum_{d \in \overline{R'}} d \tag{2.14}$$

Rocchio suggested three variables (α, β and γ) that regulate the influence of an initial query, relevant documents (positive feedback) and irrelevant documents (negative feedback), respectively. These three variables are set depending on empirical experimentation [Baeza-Yates and Ribeiro-Neto, 1999; Croft et al., 2009]. Relevance feedback techniques can be repeated multiple times [Harman, 1992] to derive a better query. For instance, Equation 2.14 is the result after one iteration. Nevertheless, collecting relevance judgments from users between iterations is time consuming, leading to investigations of the effectiveness of these approaches in the first iteration [Salton and Buckley, 1990].

We study the effectiveness of relevance feedback for ranking sentences to make query-biased summaries by adopting Rocchio's approach. While a shortcoming of relevance feedback is that it has to be manually performed by users, in our work, we assume the relevance and non-relevance of documents based on an initial ranked list of results; that is, a pseudo-relevance feedback approach. Rocchio's approach is explained in detail in Section 4.2.2.

2.6.2 Pseudo-Relevance Feedback

Pseudo-relevance feedback techniques for query expansion aim to automatically induce relevance feedback and ignore human document assessments [Croft and Harper, 1979; Xu and Croft, 1996]. Three main approaches for gathering expansion terms are: *knowledge structures*, which employ pre-constructed thesaurus-like resources and dictionaries; *global analysis*, which collects expansion terms from whole-collection statistics; and *local analysis*, which uses only the top retrieved documents by assuming that these are relevant and may share similar vocabulary. In following sections, we provide a brief review of these approaches regarding document retrieval, particularly we survey their application in query-biased summaries.

2.6.3 Knowledge Structures

Knowledge structures are linguistic resources that aim to characterise a term or relationships among terms. Efthimiadis [1996] classified such resources into three types: dictionaries, general thesauri, and domain-specific thesauri. For example, dictionaries provide definitions, and thesauri identify close relations among words such as synonymy (*similar to*) or hierarchy (*parent-child*). Consequently, terms from definitions or relations can assist in the look-up of supplementary words for query expansion.

The development of a general thesaurus such as WordNet [Miller, 1995; Fellbaum, 1998], which also acts as an electronic dictionary, enables to use more specific relations such as

Relation type	Term	Definition
Synonym	`psyche`	The immaterial part of a person; the actuating cause of an individual life
Synonym	`spirit`	The vital principle or animating force within living things
Hypernyms	`vital principle`	A hypothetical force to which the functions and qualities peculiar to living things are sometimes ascribed.
Holonym	`MEMBER OF: people`	—

Table 2.1: Three relations (synonym, hypernym and holonym) of six obtained from WordNet 3.0 for the term "soul".

hypernyms (*is a*), holonyms (*member of*), and meronyms (*part of*). For example, Table 2.1 illustrates that the word "*soul*" has multiple relations according to WordNet. As mentioned before, these terms can be applied to extend an original query. Voorhees [1994] studied different heuristics to expand query terms using WordNet. In her experiments the original query terms received higher weights, whereas words extracted from WordNet were assigned lower weights. Her experiments showed that despite selecting extra terms manually, retrieval effectiveness in large text collections was barely improved.

In contrast to general thesauri, domain-specific thesauri describe the associations of terms in a particular field of application. Nevertheless, their use is sparse given their availability. Domain-specific thesauri have been developed in areas such as Medicine [Humphreys and Lindberg, 1989], Agriculture[2] or Arts and Architecture [Petersen, 1990], to mention a few. Hersh et al. [2000] studied synonyms, hierarchical relations and definitions from the UMLS Metathesaurus[3] to improve retrieval in the medical domain. Their findings indicated that using parent-child relations and definitions to expand the query significantly deteriorated system performance, while manually selected synonyms achieved similar performance compared with unexpanded queries. Overall, research has shown that synonyms gathered from general thesauri or domain thesauri typically moderately improve document retrieval [Guo et al., 2004; Hersh et al., 2000; Voorhees, 1994].

[2] `http://aims.fao.org/standards/agrovoc/about`

[3] `http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html`

Knowledge Structures Assisting Summarisation

Generally, WordNet is used to mine synonyms of query terms in order to boost the rank of relevant sentences. However, research has shown minimal improvement in tasks associated to sentence retrieval [Zhang et al., 2004] and multi-document summarisation [Vanderwende et al., 2007], and has reported null improvement for single-document summaries [Amini et al., 2005].

Vanderwende et al. [2007] found that equivalent terms obtained from morphological variations of words are more effective for the construction of multi-document summaries than WordNet synonyms. Morphological transformations of a word are generated from changing its grammatical function to another, and preserving the same spelling; consequently, this approach does not require a word-disambiguation process. Stemming algorithms, described in Section 2.1.1, can easily produce morphological transformations, although these are not necessarily accurate. For example, the noun "`improvement`" after removing the suffix *-ment* takes the form of the verb "`improve`". However, the term `improvement` can adopt any of its multiple synonym instances such as *"advancement, betterment, change, enhancement",* or *"upgrade"*.

Word polysemy, a term having multiple meanings, is one of the main problems with knowledge structure approaches. The term `soul` in the previous example has another meaning as *"the folk music consisting of a genre of a capella music originating with Black slaves in the United States"*. That is, a word disambiguation process (automatic or manual) must be carried out to find the correct sense of a word in a given context. Another limitation of general thesauri is that relations are made by humans, hence the update process is slow [Suchanek et al., 2008]. While other more complex knowledge structures have been used in document retrieval such as ontologies [Bhogal et al., 2007], their application in summarisation methods is emerging, since generally they are restricted to a specific domain [Afantenos et al., 2005; Nenkova and McKeown, 2011].

Given the drawbacks of linguistic resources, other approaches for supporting summarisation methods can involve other external resources that do not necessarily have a linguistic foundation. These alternative methods include consulting query logs of past queries [Garg and Dhamdhere, 2012; Scholer et al., 2004; Sun et al., 2005]. For example, a document can be returned by an IR system as an answer to multiple previously submitted queries, since it may cover different topics. Thus, these query terms are assumed to be useful for the construction of generic summaries of documents [Sun et al., 2005].

In the absence of query logs, Sun et al. [2005] proposed creating lexicons from Web summaries of the Open Directory Project[4] (ODP) to locate extra terms for assembling summaries. While documents in the ODP are hierarchically organised and may serve to expand the original query, the generation of such lexicons can carry several inconveniences. We argue that ODP summaries are limited in the scope of topics they cover. In addition, lexicons require being updated, as ODP summaries are edited constantly by different contributors.

2.6.4 Global Analysis

Resources aided by linguistic foundations and handcrafted by experts has led to poor results in document retrieval [Guo et al., 2004; Hersh et al., 2000; Voorhees, 1994]. Moreover, thesauri are costly resources that require human involvement [Suchanek et al., 2008]. Global analysis techniques aim to reduce cost resources that the creation of thesauri requires. Global techniques rely on analysing whole text collections to derive statistical associations [Jing and Croft, 1994; Smeaton and van Rijsbergen, 1983] or clusters of words [Spärck-Jones, 1971] can be used to expand a query.

Statistical associations help to automatically construct thesauri or tree-like structures. An automatic thesaurus, for example, is created by finding statistical links among terms or phrases in text spans (sentences or paragraphs) from the whole text collection based on term frequencies [Jing and Croft, 1994]. In order to keep a representative amount of associations, highly frequent and rare words are ignored for the construction of such associations. In contrast to manual thesauri, automatic thesauri did improve effectiveness in large text collections [Jing and Croft, 1994].

However, global analysis techniques can harm retrieval as in the case of tree-like structures [Smeaton and van Rijsbergen, 1983], or lead to some computing limitations such as storing and updating term-to-term associations when new documents are added in the collection [Jing and Croft, 1994]. In particular, the high dimensionality of the data for large modern collections makes the initial analysis computationally expensive. Even though associations of automated structures are not attached to any domain, they are collection dependent, which may limit their re-usability for other collections.

[4] `http://www.dmoz.org/`

Global Analysis Assisting Summarisation

Little research has focused on global techniques for summarisation purposes. However, we note that clustering techniques can resemble global analysis techniques for addressing multi-document summarisation. Radev et al. [2004] obtained a set of general salient terms, called "*centroids*", after grouping documents and weighting terms with the TF*IDF approach. Centroids were used in combination with a sentence position heuristic to produce generic multi-document summaries. Another approach consists of identifying word-clusters that co-occur with the same frequency in a set of topically related documents [Amini and Usunier, 2007], similar to the concept of Local Context Analysis explained in the next section. As can be seen, global analysis techniques can seldom be found for single document query-biased summarisation. We did not explore these approaches in this thesis, given the shortcomings of gathering terms using global techniques as described in the previous section.

2.6.5 Local Analysis

Local analysis was proposed to alleviate the disadvantages of global analysis techniques. Local analysis approaches assume that only the top N retrieved results are relevant, instead of exhaustively computing statistics from entire text collections [Billerbeck, 2005; Buckley et al., 1995; Croft and Harper, 1979; Xu and Croft, 2000]. Local analysis is a parameterised approach, that is, it depends on finding an optimal number of top N documents to conduct the expansion process, and on determining the number of extra terms E to add to the original query. Billerbeck and Zobel [2003] explored the parameter space to discover optimal N and E values, and concluded that both are sensitive to the collection type. Therefore, the generalisation of such parameters among collections may not equally benefit all queries.

A negative effect of local analysis is that highly ranked documents may not be relevant to the query. If this occurs, the expanded terms are taken from irrelevant elements which can degrade retrieval, where an expanded query can derive a set of documents that are not topically related to the original request. In order to minimise such an effect, Xu and Croft [1996; 2000] proposed a technique called Local Context Analysis (LCA) to weight terms in top-ranked documents according to their co-occurrence with the original query. LCA employs *"concepts"*, which are represented as single words or noun phrases. Candidate concepts for expansion are searched in short passages or whole documents from top results, also called the local set. Xu and Croft [1996; 2000] proposed a function $f(c, Q)$ that weights a concept

c given a query Q as follows:

$$f(c, Q) = \prod_{q_i \in Q} (\delta + co\_degree(c, q_i))^{idf(q_i)} \tag{2.15}$$

A smoothing δ factor is introduced in the exceptional case that the *co_degree* component is zero. This component measures the co-occurrence degree of concept c and a query term q_i defined as:

$$co\_degree(c, w_i) = \log_{10}(co(c, q_i) + 1) \times \frac{idf(c)}{\log_{10}(R')} \tag{2.16}$$

where $co(c, q_i)$ corresponds to the number of co-occurrences of c and q_i in the local set, and $idf(c)$ represents the inverse frequency of concept c in the collection normalised by the number of top results R'. The following equations show the definition of $co(c, q_i)$ and $idf(c)$:

$$co(c, q_i) = \sum_{d \in D} f(c, d) \times f(q_i, d) \tag{2.17}$$

$$idf(c) = min(1.0, \log_{10}(N/N_c)/5.0) \tag{2.18}$$

The frequency of c and q_i in document d is given by $f(c, d)$ and $f(q_i, d)$, respectively. Finally, N is the total number of documents in the collection and N_c is the number of documents where concept c appears.

Xu and Croft [2000] tested the effectiveness of LCA in several text collections and different query styles. They found up to 23% increase using the TREC-3 and TREC-4 collections, where query descriptions are short. However, LCA showed modest but not significant improvements in cases where the queries of the collection were good descriptors of the information need.

Local Analysis Assisting Summarisation

We observe that knowledge structures and global analysis approaches can be suitable to create summaries for specific audiences and; therefore, to particular domains. However, our research focuses on generating query-biased summaries for general audiences. Given shortcomings of knowledge resources and global analysis approaches, we investigate local techniques to construct query-biased summaries of documents towards informational requests. Usually IR systems do not group documents by their topical similarities; rather documents are returned and ranked according to their resemblance to a submitted query. Hence, we suggest that

highly ranked documents share a certain vocabulary related to the initial query. Based on former research, we identify two main trends of local analysis that have been explored for improving summarisation techniques or sentence retrieval tasks. We identify that local analysis can be carried out at the *sentence level* or *document level*, both explained below.

Local Analysis at Sentence Level. This approach consists of finding expansion terms in a set of top ranked sentences that come from top ranked documents. Consequently, the search space to locate extra terms is reduced to smaller passages, in this case sentences. Using the TREC Novelty track dataset, local analysis at the sentence level has been investigated for passage retrieval tasks for example by Losada [2010]. Words in top-ranked sentences can be used in multiple ways to source expansion terms, these approaches include: to conduct a part-of-speech process for identifying nouns as extra terms; to employ term-weighting techniques [Ko et al., 2008]; and to find query-term co-occurrences similar to the LCA approach [Amini et al., 2005; Losada, 2010].

While Han et al. [2000] and Ko et al. [2008] reported that sentence-level approaches are effective, Goldstein et al. [1999] found that using words from the document title improved against using terms of the first ranked sentence of a document. The performance of sentence level methods is unclear, since experiments were conducted on collections that lack formal topic descriptions or in small testing settings [Amini et al., 2005; Han et al., 2000]. For example, the title of documents were employed to mimic users requests to locate a set of top sentences for the expansion process. Amini et al. [2005] found that expanding the title of documents with clusters of words co-occurring in sentences was effective but not robust among collections of patents and scientific articles.

Local analysis at the sentence-level may carry additional efficiency concerns. After retrieving the top documents, a second pass is required to rank the sentences that will be employed for mining expansion terms [Goldstein et al., 1999; Han et al., 2000; Ko et al., 2008; Losada, 2010]. This may require extra disk space or increase the response time [Ko et al., 2008]. We did not explore these efficiency shortcomings in our work. By employing a large testbed such as the TREC Novelty track, described in Section 2.9.1, we examine whether sentence-based expansion approaches are effective for ranking sentences that will compose query-biased summaries.

Local Analysis at Document Level. This approach is comparable to the task of query expansion for document retrieval, where an initial set of top-ranked documents is used to

source extra terms, and no additional sentence retrieval is needed. Sanderson [1998] studied document-based expansion to improve passage selection of documents related to a specific topic. His results indicated that chosen passages often drifted topically to other content, rather than focusing on information expressed in initial requests. However, the original queries were notably detailed, approximately 38 words excluding stopwords. Consequently, an expansion of 70 terms did not improve over the original queries. These findings support the idea that query expansion can be helpful in contexts where the request is short and ill-constructed [Xu and Croft, 2000], which can be applicable to the generation of query-biased summaries. Local analysis at document level has been shown to improve sentence ranking from multiple documents related to the same topic for passage retrieval tasks. For instance, Losada [2010] applied LCA to complement the original query to increase the selection of sentences judged as relevant. LCA showed robust results among the TREC Novelty track 2003 and 2004 datasets for finding a set of relevant sentences from multiple documents related to the same information request [Losada, 2010]. However, this was not investigated for query-biased summaries as we aim in this work.

As can be seen, there is relatively little research in terms of single-document query-biased summaries assisted by query expansion. While the approaches of Losada [2010] and Sanderson [1998] are close to our work, there is a gap to investigate the effects of the expansion in sentence ranking methods for assembling short query-biased summaries. We present results for document-based and sentence-based expansion approaches in Sections 4.4.2 and 4.4.3, respectively.

2.6.6 Other Query Expansion Techniques

Other sophisticated expansion approaches have come up along side relevance feedback and pseudo-relevance feedback techniques. For example, interactive query expansion provides users with terms that can be related to their requests in the form of suggestions. While users are in charge of selecting closer terms to extend the query, research has demonstrated that users do not necessarily choose the best terms to expand their query [Ruthven, 2003]. Moreover, gathering these recommendation terms also relies on automatic methods such as those explained in Section 2.6.2.

Buscher et al. [2008] employed eye tracking techniques to record gaze patterns of users to identify possible areas of interest in a document. Hence, terms located in these areas were employed to extend the query. Despite knowing implicit preferences of users in real time, the

approach is expensive and limited to laboratory settings, hence we do not pursue it in this thesis. As future work an exploratory study can investigate if automatic query expansion techniques analysed in this dissertation (Rocchio's approach or LCA) identify the same terms that users focused on while reading as identified by eye tracking.

2.7 Development of Testbeds

Testbeds are important for evaluating the performance of IR systems and for replicating results for further comparison. In a document retrieval scenario, a testbed consists of a set of topics, documents and relevance judgments [Cleverdon, 1967; Kelly, 2009]. The assembly of testbeds is often delegated to specialised organizations or group-based evaluation exercises given the human effort and cost involved in their construction. The National Institute of Standards and Technology (NIST) under the Text REtrieval Conference[5] initiative (TREC) has developed testbeds for investigating document retrieval in large corpora. Generally, TREC invites the research community to participate in information retrieval challenges (tracks), and provides the infrastructure for experimentation and evaluation. These tracks usually explore current problems in areas of IR.

However, TREC has not focused on studying summarisation. Other NIST and independent initiatives have created specific tracks for it. In 2008, NIST created the Text Analysis Conference[6] (TAC) that aims to study Natural Language Processing problems on a large scale. TAC, known as the Document Understanding Conference[7] (DUC) from 2001 to 2007, has proposed tracks to investigate different summarisation styles, as well as multiple summary intrinsic evaluation methodologies. Nevertheless, most of the DUC/TAC conferences have studied multi-document summarisation as can be seen in Table B.1 in Appendix B. In 2011, the Initiative for the Evaluation of XML retrieval[8] (INEX) introduced the Snippet Retrieval track. This track explores not only retrieval document approaches, but also the generation of short summaries that can provide information about of a document at a glance [Trappett et al., 2011].

Despite the fact that TREC and DUC/TAC have slightly different aims, both may share elements of testbeds such as topics, documents and relevance assessments. We continue describing typical elements in testbeds; and, where pertinent we provide details of such

[5] `http://trec.nist.gov/`
[6] `http://www.nist.gov/tac/about/index.html`
[7] `http://www-nlpir.nist.gov/projects/duc/index.html`
[8] `https://inex.mmci.uni-saarland.de/`

Topic:	353
Title:	Antarctica exploration
Description:	Identify systematic explorations and scientific investigations of Antarctica, current or planned.
Narrative:	Documents discussing the following issues are relevant: systematic explorations and scientific investigations of Antarctica (e.g., seismology, ionospheric physics, possible economic development); other research currently conducted or planned for the future; banning of mineral mining. Documents discussing tourism are non-relevant. Documents discussing "disrupting scientific experiments" are non-relevant unless a specific experiment is identified.

Table 2.2: Components of a TREC topic.

components regarding summarisation tasks. We also explain model summaries, which are a key component in summary evaluation.

2.7.1 Topics

A topic is a set of statements that models the scope of an information need, which can be answered using documents of text collections. Topics are intended to resemble real information requests, and are typically formulated by experts. Usually, topics include several fields: a numeric ID; a title, which is a succinct list of key terms about the topic; a description, which is a brief explanation of the topic; and a narrative, which specifies multiple facets of the topic and limits the information need. For example, an information request about *explorations in Antarctica* prepared by an expert is shown in Table 2.2. The availability of topics in testbeds plays an important role, since the title, the description and the narrative can serve as baseline requests for evaluation purposes.

2.7.2 Documents

Documents in text collections may vary in terms of length, vocabulary, structure, language and formats, to mention a few. Documents can be gathered from: newswire services [Dang, 2005], congressional records,[9] Web pages from the `.gov` domain [Clarke et al., 2004], general Web content,[10] set of patents [Amini et al., 2005], research articles [Kupiec et al., 1995], and Wikipedia articles [Trappett et al., 2011]. For instance, the AQUAINT corpus is a compilation of articles from different newswire services and has been employed in TAC

[9] `http://trec.nist.gov/data/docs_eng.html`
[10] `http://lemurproject.org/clueweb09/`

conferences [Dang and Owczarzak, 2008] and TREC tracks [Harman, 2002; Soboroff and Harman, 2003; Soboroff, 2004]. Articles in newspapers can follow events such as natural disasters, accidents, political campaigns or controversial topics in a specific time frame. This kind of document has enabled the investigation of issues such as paraphrasing [Gaizauskas et al., 2001] and dealing with redundant content in multi-document summaries [Dang and Owczarzak, 2008].

2.7.3 Relevance and Relevance Judgements

Before explaining relevance judgements, we have to explain the concept of *relevance*. Saracevic [2007a;b] pointed out that relevance is an "elusive" and "timeless" concept to describe, with no precise definition among the IR community [Borlund, 2003; Mizzaro, 1997; Saracevic, 1975; 2007a;b]. Mizzaro [1997] proposed that relevance is a relation between two entities: a document (or surrogate), a physical representation of information; and a problem of requiring information (also information need or request). However, his definition extends the observation that relevance is shaped by multiple factors such as topicality, users, type of judgement, task, or temporary, for mentioning some.

Topical relevance requires a document to contain similar content to an information request. For instance, a document that discusses *scientific explorations in Antarctica* can be topically related to the request of marine wildlife in Antarctica or research vessels in the Southern hemisphere. User relevance deals with subjectivity; what is relevant to one person might not have the same relevance value for someone else. In order to describe how relevant a document is, users can assign a value relying on a binary scale (yes/no) or a multi-scale (non-relevant, barely relevant, relevant, highly relevant) approach [Cleverdon, 1967], for example. However, the assessment can vary to different levels of relevance. In addition, as assessors progress in a judging task, they become more confident or familiar with the topic, which can relax their judgements. So what seemed to be relevant at the beginning of a task may not be after a while [Saracevic, 2007a; Scholer et al., 2011].

The assessments where assessors judged the relevance of documents are known as *relevance judgements*. Relevance judgements are valuable because joined with topics and documents they form a framework for evaluation. While there are criticisms regarding the nature of relevance judgements [Saracevic, 2007b], we need to recall that they merely gauge topical relevance of a system [Borlund, 2003].

Relevance judgements can be given not only for documents, but also for document rep-

resentations [Barry, 1998; Janes, 1991; Marcus et al., 1978], or sentences [Harman, 2002; Soboroff and Harman, 2003; Soboroff, 2004]. Document representations — also called as surrogates by Mizzaro [1997] — are descriptors of a document that can be generated from its content or done by an assessor. For instance, a document representation can include any of the following descriptors: a title, key terms, authors, type, an extract, an abstract (given by the author or a librarian), to mention a few. In the case of sentence judgements, an assessor can be asked to identify sentences that are relevant to a specific topic as in the TREC Novelty track. We detail this track in Section 2.9.1.

2.7.4 Model Summaries

In addition to reliable testbeds, summary evaluation may involve a set of model summaries — also known as *gold standard, ground truth, ideal summaries* or *reference summaries*. This set is "ideal", since the summaries are the product of an intellectual process performed by a human. Hence, an automatic summary should aspire to have certain similarity with its ideal counterpart.

There are several ways to compile model summaries. In a simple approach, an assessor creates a model summary of a document by selecting sentences that are deemed important and then these sentences are concatenated [Edmundson, 1969; Goldstein et al., 1999; Jing et al., 1998; Marcu, 1997; Rath et al., 1961]. This process may resemble the collection of relevance judgments at the sentence level, as explained above. In a more sophisticated approach, model summaries can be the product of a cognitive task, where assessors are asked to compose an abstract; thus they produce new prose [Halteren and Teufel, 2003]. Abstracts require subjects to employ a deeper understanding of the original document by merging several ideas into one single sentence, paraphrasing or generalising content [Brown and Day, 1983; Irwin and Doyle, 1992]. The collection of abstracts as ground truth is common in DUC/TAC evaluations as can be seen in Table B.1 in Appendix B. The expertise of an assessor is valuable as it helps in creating reliable and stable model summaries. However, in a less rigorous fashion, subjects with good literacy skills can be recruited to collect sentence judgements or to compose model summaries.

In the absence of assessors, previous research has used answers from "Frequently Asked Questions" pages. That is, the answer provides a succinct description given a request [Berger and Mittal, 2000]. Another alternative approach is to employ the abstracts written by the author of the document; this depends on the type of documents such as scientific articles. The

content of these abstracts can then manually be mapped with the original document to get an aligned pseudo-gold standard of sentences [Amini et al., 2005; Kupiec et al., 1995]. Given the laborious task of finding sentences in the document that match in meaning with a summary sentence, automatic alignment methods have been proposed. Alignment algorithms can trace original sentences [Marcu, 1999] or specific fragments of text [Jing and McKeown, 1999] where a summary sentence comes from. However, a trade-off is involved since the alignment may involve certain margin of error with respect to the original abstract by introducing unrelated sentences.

Given our approach to rank sentences for constructing query-biased summaries, we use topics, documents and relevance judgements of sentences from the TREC Novelty track to gauge the performance of sentence ranking methods. More details regarding this track are provided in Section 2.9.1. In the next section, we examine existing intrinsic and extrinsic methodologies for summary evaluation.

2.8 Summary Evaluation Approaches

The evaluation of automatic summaries can be conducted using *intrinsic* or *extrinsic* methodologies. Intrinsic methodologies quantitatively gauge the content of summaries by measuring the overlap of words or sentences in both automatic and model summaries. Extrinsic methodologies, on the other hand, qualitatively assess summaries by exposing them to the scrutiny of users in the process of performing specific tasks. However, Spärck-Jones [2007] suggested that this classification was too broad, and defined two levels of evaluation in both methodologies, as shown in Figure 2.2. Intrinsic methodologies can be *quasi-* or *semi*-purpose. According to this classification, intrinsic quasi-purpose evaluations can rely on model summaries or sentence relevance assessments to compare automatic summaries. We follow this methodology in Chapters 4 and 5. The framework for a quasi-purpose intrinsic evaluation may involve two key elements: a testbed and a set of model summaries, explained in Section 2.7. Intrinsic semi-purpose methodologies inspect linguistic features of summaries such as language appropriateness, readability, or coherence. This approach is out of the scope of this thesis, since summarisation methods we study excerpt sentences from a document rather than composing new prose.

Extrinsic evaluation involves *pseudo-* or *full*-purpose methodologies. In pseudo-extrinsic methodologies, subjects assess summaries whether these are useful, informative or indicative (depending on summary functionality) in a given simulated context. We present results of a

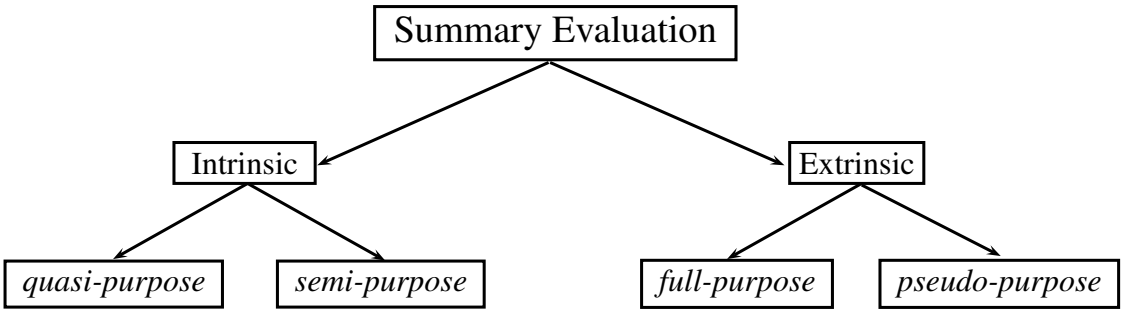

Figure 2.2: Summary evaluation classification.

pseudo-extrinsic approach in Chapter 6. In contrast, full-purpose methodologies provide not only a simulated context, but also asks users complete particular tasks with the assistance of summaries. In Section 6.2, we explain the reasons of not using full-purpose methodologies in our work. Section 2.8.2 briefly surveys the characteristics of extrinsic evaluations.

2.8.1 Intrinsic Evaluation

Intrinsic methodologies assess summaries by measuring the overlap of units (words or sentences) between automatic and model summaries. The overlap is expressed in the form of a score or proportion that indicates the coverage of an automatic summary relative to a model summary. The popularity of intrinsic methodologies is due to their nature of *black boxes*, since the input is a set of model and automatic summaries, and the output is a score showing how well a summarisation method has achieved. That is, the higher the overlap, the better the summary. We identify two main methods to conduct intrinsic evaluation: string and content matching. In this thesis, we employ a string matching approach, since it is less human-dependent. These approaches are explained in the following sections.

String Matching Methods

Intrinsic summary evaluation can use existing measures of document retrieval such as precision, recall and F1-score [Goldstein et al., 1999; McKeown et al., 2005]. These measures can be modified to assess the overlap of information units such as either sentences, words or sequences of words between automatic and model summaries. Precision is defined as the ratio between number of units in both model (M) and automatic (A) summaries divided by

the number of units in the automatic summary. Precision is computed as:

$$precision = \frac{|M \bigcap A|}{|A|} \tag{2.19}$$

Recall represents the ratio between the number of content units in both model and automatic summaries, divided by the number of units in a model summary. Consequently, recall is defined as:

$$recall = \frac{|M \bigcap A|}{|M|} \tag{2.20}$$

Finally the F1-score is the harmonic mean of precision and recall as shown below:

$$F1 = 2 \cdot \frac{precision \times recall}{precision + recall} \tag{2.21}$$

We employ $P@n$ for gauging summarisation methods, a measure to assess the effectiveness of ranked document retrieval. $P@n$ is based on the supposition that in large text collections users are not able to inspect all results, rather they only assess some of the top returned documents [Büttcher et al., 2010]. Similarly, this can be applied for sentences in documents, where only the n best scored sentences are appealing for constructing a summary. In this case a model summary M contains a list candidate sentences that can be included in the summary. We provide results of the evaluation using this approach in Chapters 4 and 5. $P@n$ is determined as:

$$P@n = \frac{|M \bigcap A[1 \dots n]|}{|A[1 \dots n]|} \tag{2.22}$$

Automatic evaluation in early DUC conferences calculated the proportion of machine summaries that appeared in model summaries using precision and recall. In 2004, the ROUGE package (Recall-Oriented Understudy for Gisting Evaluation) was introduced to provide a formal and uniform tool for the evaluation of DUC participants' systems. ROUGE measures the similarity between multiple model summaries and automatic summaries by comparing n-grams and word sequences [Lin, 2004]. ROUGE-n indicates recall values by using word arrangements of length n. Word sequences, in contrast, consist of skip-bigrams that are unigram pairs separated by up to four words. Typical ROUGE measures applied in DUC/TAC conferences were ROUGE-1, ROUGE-2 and ROUGE-SU4 because were shown to correlate well with human judgements [Lin, 2004]. The main critique of ROUGE was based on its nature of computing similarity in terms of vocabulary matches in the form of n-grams and word sequences, instead of the informative content of the summary. Despite strong accep-

tance of ROUGE by the summarisation research community in general, other methods such as the Pyramid method [Nenkova and Passonneau, 2004] and Basic Elements [Hovy et al., 2006] have been introduced in summarisation task of DUC/TAC conferences, see Table B.1 in Appendix B.

Content Matching Methods

In contrast to string matching methods, other intrinsic methodologies aim to assess summaries based on semantic units of content. Content matching methods acknowledge that one model summary is insufficient to evaluate automatic approaches, as there is not a definite human summary that encompasses a unique ground truth [Nenkova and Passonneau, 2004]. These methods require further manual inspection of multiple model summaries from the same document to identify *content units*. Content units of model summaries can take the form of factoids [Halteren and Teufel, 2003] or Summary Content Units (SCUs) [Nenkova and Passonneau, 2004].

Factoids are short statements that are syntactically different, but that convey the same meaning. Table 2.3 shows an example of factoids found in model summaries. This approach is costly as a large number of model summaries (from 30 to 40) are required to find consensus of factoids. When more model summaries are added, the list of factoids has to be updated. Another shortcoming of this approach is that assessors have to make clear distinctions whether sentences convey opinions from authors of model summaries, or factual events [Halteren and Teufel, 2003].

SCUs are key elements of the Pyramid method, where assessors firstly identified similar sentences in different model summaries. On close examination, assessors have to detect nuggets of information that express the same idea. These nuggets are formally called SCUs, which receive a weight depending on the number of summaries they appear. SCU are accommodated in layers (a pyramid), where top layers correspond to highly frequent SCUs, and vice-versa for those with low appearance in model summaries. The pyramid score for an automatic summary is given by the ratio of the sum of SCUs weights in the summary, and the optimal summary content (includes all SCUs in the top layer). Table 2.3 also illustrates an example of SCUs obtained from two model summaries. Despite being able to evaluate summary content more accurately, assessors require intensive training to obtain a reliable set of content units [Nenkova and Passonneau, 2004]. We did not conduct content matching evaluation, as our approach focuses on string matching methods at the sentence level,

Factoids	
Document Sentence:	*The pandas' main sanctuary, Sichuan, is China's most populous province with 100 million people, where animals have an uphill battle to survive.*
Factoid 1:	`Pandas are endangered.`
Factoid 2:	`The approach of man menaces the existence of pandas.`
SCUs	
Similar sentences in model summaries:	
Sentence 1:	*Sichuan is the most populous province in China threatening panda's habitat.*
Sentence 2:	*Pandas live in densely populated areas of China.*

Table 2.3: Example of factoids and SCUs. SCUs have been undelined for easy identification.

detailed in Section 4.3.

Other Intrinsic Methods

In contrast to content matching approaches, Radev and Tam [2003] proposed a more flexible mechanism to judge summaries, which consists of measuring the utility of sentences in a document. That is, sentences in a document are judged on a scale from one to ten according to the overall informative load they may contribute in a summary. The oracle utility is calculated by averaging the scores of the most popular sentences in all judges' assessments. Thus, a summary is evaluated in terms of its *relative utility* with respect to the oracle. However, the approach involves a more demanding task due to sentence assessments not being made on a binary scale. Another intrinsic approach suggests to compute the cosine similarity between summaries and documents in the absence of a gold standard set [Donaway et al., 2000]. While the approach is appealing, it has the inconvenience that the cosine similarity treats terms independently. From 2009 to 2011 the TAC conference ran the Automatically Evaluating Summaries of Peers (AESOP) task. Its aim was to develop new metrics for assessing summaries; however, no further efforts have been developed.

2.8.2 Extrinsic Evaluation

Intrinsic evaluations particularly focus on quantitative aspects of automatic summaries; however, assessors can judge a summary based on many other criteria such as non-redundancy, overall responsiveness, readability, fluency, structure or coherence. For example, overall re-

sponsiveness and readability are some criteria used in DUC/TAC conferences as shown in Appendix B. Unfortunately, there are no automatic techniques that support research in this regard. Extrinsic methodologies expose summaries to users in the context of performing specific tasks. The summary is not gauged in terms of sentence or vocabulary matching, but rather on how supportive it is to guide users in their endeavours. Typical extrinsic evaluations are conducted in a laboratory setting aiming to reproduce an every-day enterprise (specific activity), where summaries assist subjects to complete given labours. In an experimental scenario, subjects can be given summaries to perform tasks such as: obtaining updated information from newspapers articles [Dang and Owczarzak, 2008]; finding useful results in large text collections [Tombros and Sanderson, 1998; White et al., 2003], or discovering information within a document [Egan et al., 1989]. During the study, the performance of participants employing or interacting with the summary is under analysis, rather than the participants themselves.

The first formal framework of extrinsic evaluation was proposed by Hand [1997], which subsequently was achieved by the Defense Advanced Research Projects Agency in the TIPSTER Text Summarisation Evaluation (SUMMAC). The SUMMAC conference was the former initiative to conduct a large extrinsic evaluation of automatic summaries [Mani et al., 1999]. The evaluation included many of the aspects proposed by Hand [1997], particularly in terms of the tasks such as ad-hoc retrieval, categorisation and question answering. Useful results from this assessing effort were obtained regarding summaries for ad-hoc retrieval tasks, showing that assessors made quicker judgments while identifying relevant documents given the summaries. To the best of our knowledge, formal efforts of this magnitude for extrinsic evaluation have not been attempted since then. This is most likely due to the amount of human effort that is required. We present outcomes of an extrinsic evaluation approach in Chapter 6, where subjects are asked to select the summary that would be more helpful in order to decide whether to click, and further read an underlying document.

2.9 Evaluating Sentence Ranking Methods

Methodologies developed by DUC/TAC conferences assessed summaries in terms of vocabulary overlap [Lin, 2004] or content matching units [Nenkova and Passonneau, 2004] against a set of model summaries. This set is comprised of abstracts authored by assessors, who may merge ideas into a single sentence or may paraphrase content. Consequently, these model summaries can make more complex the evaluation of automatic sentence ranking approaches.

Moreover, most of the DUC/TAC conferences investigate multi-document summarisation (see Appendix B), but not single-document. The INEX Snippet Retrieval track, on the other hand, only carried out an extrinsic evaluation to identify potential relevant Wikipedia articles. However, it lacks of model summaries in order to conduct further intrinsic evaluations [Trappett et al., 2011]. Hence, testbeds provided by DUC/TAC conferences or the Snippet Retrieval track cannot be used straightforwardly to quantitatively assess sentence ranking methods such as those studied in this thesis.

In a search engine setting that employs the Web as a corpus, documents can include a wide variety of formats and content. For example, news articles can be embedded in Web pages; however, the content and writing style they use might be more formal than Web sources such as those found in forums or blogs. Little research has been conducted to identify relevant sentences of Web content in TREC tracks [Wang et al., 2007], or using the Web as a corpus [Ko et al., 2008]. The latter approach may not enable replication of experiments, given that the Web is a continuous changing repository [Kelly, 2009]. Moreover, sentence relevance assessments of Web content usually include only a small number of queries or documents [Ko et al., 2008; Varlamis and Stamou, 2009; Wang et al., 2007].

While TREC has included corpora other than news articles, TREC does not provide tracks for studying summarisation. However, some of these tracks such as the Novelty track can be adapted to assess snippets [Metzler and Kanungo, 2008]. We mentioned in Section 2.7.3 that relevance can be collected not only for documents, but also for sentences such as in the TREC Novelty track [Harman, 2002; Soboroff and Harman, 2003; Soboroff, 2004]. This track is a compilation of news articles from the AQUAINT corpus. Sentence assessments of this track are appealing for evaluating sentence ranking approaches from sources that may help users satisfy their informational requests, which is the type of request under study in this work. In the next section, we detail the TREC Novelty track as this is the collection we employ in our experiments.

2.9.1 The TREC Novelty Track

TREC ran a Novelty track from 2002 to 2004 [Harman, 2002; Soboroff and Harman, 2003; Soboroff, 2004], which aimed to study filtering methods to identify relevant and novel information. The track focused on determining relevant and novel content at the sentence level according to a specific topic, rather than retrieving documents. The TREC Novelty

track[11] made available relevance and novel sentence assessments separately. Track organisers decided to use sentences, as they can be easily identified and can reduce the complexity for discovering redundant information [Harman, 2002].

In the first year of the track, assessors had several restrictions that led to identify only a small proportion of sentences as relevant (2%). These restrictions involved: judging sentences independently and not selecting contiguous sentences; not creating topics; and proving sentence judgements based on a short topic description. Organisers have suggested that the outcomes of the Novelty track 2002 should only be viewed as a pilot study [Harman, 2002]. Hence, we only employ data from the Novelty tracks 2003 and 2004 in our experiments.

The Novelty tracks 2003 and 2004 are homogeneous in terms of constructing topics and gathering judgements [Soboroff and Harman, 2003; Soboroff, 2004]. NIST assessors created 50 topics from the AQUAINT newswire collection for each year of the track. The AQUAINT collection compiles articles from three newswire services in specific time frames such as the Associated Press (1998-2000), the New York Times News Service (June 1998-September 2000) and the English version of the Xinhua News Service (January 1996-September 2000). Topics derived from this collection were classified either as "Opinion" or "Events". Opinion topics contain documents that describe the course of a polemic matter, while Event topics depict a natural disaster or political conflicts, for example.

In addition to topic construction, assessors had to identify relevant documents for the topic requests they created. For this task, they employed the WebPRISE IR system. The Novelty track 2003 was comprised of 25 relevant documents per topic. However, some irrelevant documents were included in the Novelty track 2004 data. These irrelevant documents were below the first 25 ranked results, and were added intentionally in order to increase the complexity of this year's task for participants. However, assessors only judged sentences in relevant documents [Soboroff, 2004]. Table 2.4 outlines the main features for the three running years of the Novelty track.

Documents in the AQUAINT collection contain an `id`, which corresponds to the date of authorship. Relevant documents were sorted according to this identifier and split into sentences. All document sentences were pooled in one single "document" for judgement, so assessors inspected document sentences chronologically. According to a given topic, assessors distinguished sentences that they deem relevant. Moreover, they did not have restrictions to select a specific number of sentences per topic or per document as relevant. Each topic was examined by two assessors; however, only sentence judgements from the author's topic were

[11] Hereafter called as the Novelty track.

Feature	Novelty track 2002	Novelty track 2003	Novelty track 2004
Collection	TREC 6, 7 and 8	AQUAINT	AQUAINT
Topics	50	50	50
Documents per topic	25	25	≥ 25
Relevant documents per topic	25	25	25
Order of documents	Rank retrieved	Chronological	Chronological
Type of topics	Not defined	Opinions and Events	Opinions and Events

Table 2.4: Features of the TREC Novelty track: text collections, topics and judged documents in the Novelty track from 2002 to 2004.

Feature	Novelty track 2003	Novelty track 2004
Topics	1-50	51-100
Documents per topic	25	≥ 25
Documents	1,250	1,808
Sentences	39,820	52,447
Relevant sentences	15,557	8,343
Non-relevant sentences	24,263	44,104

Table 2.5: Composition of the Novelty track collections between 2003 and 2004.

taken as ground truth in the Novelty track evaluation. We also use the assessments made by the author's topic in our evaluation approach. Table 2.5 provides document and sentence statistics for the Novelty track 2003 and 2004.

We discussed earlier in this section that testbeds developed by DUC/TAC or INEX initiatives can not be used directly to gauge sentence ranking methods. We propose that sentence ranking methods can adopt not only document retrieval techniques to generate query-biased summaries, but also employ similar methodologies to evaluate the effectiveness of sentence ranking methods. By using topics, documents and sentence relevance judgements of the Novelty track, we adapt the Cranfield methodology to assist in the evaluation of sentence ranking methods. In Chapters 4 and 5, we examine the effectiveness of sentence ranking methods following this evaluation approach.

2.10 Collecting User Data

The Cranfield methodology [Cleverdon, 1967] is a framework that quantitatively evaluates the effectiveness of retrieval systems. In this scenario, human expertise is generally required to assemble testbeds as explained in Section 2.7. However, non-specialised human intervention can be required to collect data not only for evaluation, but also for experimentation.

That is, a researcher may be interested in understanding interactions between subjects and retrieval systems, which is difficult to capture through testbeds. Different techniques have been employed to obtain or to gather information from users' experiences with retrieval systems such as surveys, query logs, interviews, thinking-aloud protocols, eye tracking, or crowdsourcing, to mention a few. In this section, we only explain surveys, eye-tracking and crowdsourcing techniques as these were employed in our work to gather data. For a general overview and comparison of other mechanisms to collect data from subjects can be found in Kelly [2009].

2.10.1 Surveys

Surveys are probably the most popular and simplest way to gather feedback about perceptions and opinions of subjects. These can be administered on paper, while more sophisticated methods distribute them through email or made them available on the Web through specialised software tools such as SurveyMonkey,[12] Qualtrics[13] or GoogleDocs.[14] Questions in a survey can vary depending on the objectives of the researchers [Kelly, 2009]. For example, open questions allow investigation of users' behaviours, and closed questions (multiple choice or likert-type scales) measure quantitative aspects of a systems. Appendix D shows a survey administered in our study that we detail in the next chapter. The survey collected demographic information from participants and general experience with search engines.

2.10.2 Eye Tracking Techniques

Eye tracking techniques consists of recording gaze activity (eye movements) over a stimulus area, generally through specialised hardware and software [Duchowski, 2007]. Eye movements are characterised by *"fixations"* and *"saccades"*. Fixations are stable gazes from 200ms to 300ms, while saccades are extremely quick jumps between fixations with a maximum length of 7-8 letters [Rayner, 1998]. The size of fixations and their distribution vary depending on the stimuli, or task presented to subjects such as reading text, reading music, typing or visual search.

Eye tracker devices can be found in several forms such as mounted in glasses, helmets, or have the appearance of a typical computer screen [Duchowski, 2007]. These devices rely on optical sensors to detect gaze activity, as well as image processing and mathematical models

[12] `http://www.surveymonkey.com/`
[13] `https://www.qualtrics.com/`
[14] `docs.google.com/`

to estimate eye positions given reflection patterns of the eyes. Besides specific hardware, eye trackers provide software that summarises gaze activity over a stimuli through gazeplots or heatmaps. A gazeplot is a graphical representation of the path of fixations. Fixations are depicted as circles that enclose a set of gazes at a specific time (measured in miliseconds) and location, whereas saccades are represented as lines connecting a fixation with its corresponding previous and next fixations. Multiple overlapping circles indicate that a user fixated at different time episodes in the same stimulus area. On the other hand, heatmaps summarise the regions of a stimulus where the eye tracker registered gaze activity by colouring these areas. For instance, warm colors, in the gamma of red-orange-yellow, indicate that a specific area has registered intense eye movement activity. Cold colors, in the gamma of green-blue-purple, represent a low gaze activity.

Duchowski [2002] surveyed multiple applications of eye tracking techniques as a supportive technology in areas such as Psychology, Marketing or Human Computer Interaction. For example, former research has found that readers tend to fixate for a longer time when experiencing problems to understand a word and reading patterns [Rayner, 1998; Reichle et al., 2006], skim through text that might not be relevant [Nielsen, 2006] or spoken language comprehension problems [Liversedge and Findlay, 2000], to mention a few. Recording eye movements of users while reading search engine result pages is a typical application of eye tracking technologies in IR systems [Cutrell and Guan, 2007; Joachims et al., 2005; Lorigo et al., 2008].

An advantage of eye tracking techniques is that these enable to the investigation of human behavior given the underlying cognitive process of eye movements. Interactions between systems and users can not be explicitly gathered through log files or surveys. Thus, eye tracking can be considered as an unobtrusive technique for collecting interactions between users and systems in real time with minimal or no contact with the experimenter. We apply eye tracking techniques to track more accurately text fragments subjects employ while constructing query-biased summaries. These query-biased summaries are the product of a cognitive process difficult to trace just with a typed text. In Section 3.3, we describe the eye tracker device we employed in our user study and operational details.

2.10.3 Crowdsourcing and CrowdFlower

The final tool we use to evaluate sentence ranking methods to construct query-biased summaries is crowdsourcing techniques. Crowdsourcing is a mechanism to distribute *micro tasks*,

which are work items broken into small units. These tasks are made available to a large population of workers, who carry them out for a small payment. The tasks require human intelligence to be completed, since they cannot be straightforwardly automated. The proliferation of crowdsourcing relies on collecting large amount of completed tasks in a short period of time. In addition, the cost is relatively low compared with recruiting research participants for a laboratory-based experiment.

The crowdsourcing working schema includes two entities: requesters and workers. Requesters and workers have to register in a crowdsourcing platform to have access to the services. Requesters submit micro tasks and provide instructions to guide workers for completing them. The crowdsourcing platform lists the available tasks (after requesters have ordered them), and workers voluntarily select those they would like to perform. The complexity of a task can vary, so workers might choose tasks that seem interesting or easy to them. Although there is no restriction on time, workers may take between one or two minutes to perform a set of units depending on the complexity of the task. Typical examples of tasks performed in crowdsourcing platforms include: making relevance assessments [Alonso and Mizzaro, 2009]; rating the similarity between words [Snow et al., 2008]; or ranking the quality of translation systems [Callison-Burch, 2009], to name a few.

There are several platforms that manage operations/transactions between requesters and workers under the crowdsourcing scheme such as Amazon Mechanical Turk[15] or CrowdFlower.[16] In Chapters 5 and 6, we provide details of two user studies using the CrowdFlower platform.

Given the lack of contact between the experimenter and subjects (requester and workers), collecting information via crowdsourcing may have certain shortcomings. For example, subjects may not understand the task, lose interest because tasks are long or complex, or be spammers. CrowdFlower has a mechanism to detect workers who may not be carrying out tasks diligently. This consists of randomly presenting units with a unique and definite answer, known as *gold units*. Gold units, which are prepared by the experimenter, have the appearance of normal units, so workers can not easily differentiate them from task units. If workers fail to correctly respond a certain number of gold units, they are labelled as "untrusted", and their assessments are not included in the final pool of results. CrowdFlower suggests to employ as gold units around 5% and 10% additional to the total number of work units, and considers trusted workers those achieving more than 70% of gold units correctly

[15] `https://www.mturk.com/mturk/welcome`
[16] `https://crowdflower.com/`

answered. We followed these recommendation practices to set up our experiments detailed in Section 5.1.1 and Section 6.2.

2.11 Summary

This chapter has reviewed common practices to create query-biased summaries. In particular, we have oriented it as a sentence ranking method where sentences constitute the units of extraction. As user requests may be vague, short or ambiguous this may reduce the capability of sentence ranking methods to find good candidate sentences for a query-biased summary. That is, in document retrieval a word can be found in several areas of a text and different documents, but repeated words are less likely to occur in sentences. The next chapter presents the design and results of an exploratory study, which looks into the way humans create short query-biased summaries. This user study is a preamble to detect areas of improvement in query-biased summaries. We have surveyed query expansion techniques as a mechanism to reduce the vocabulary gap, and to boost the selection of relevant sentences. This is investigated in Chapter 4.

This chapter has also reviewed intrinsic and extrinsic methodologies to evaluate summarisation methods. We observe that evaluation approaches developed by the DUC/TAC conferences or the INEX Snippet Retrieval track do not focus on assessing sentence ranking methods. The TREC Novelty track made available relevance sentence assessments, which we employed to quantitatively gauge sentence ranking approaches. This is examined further in Chapters 4 and 5. We also conduct an extrinsic evaluation of query-biased summaries, which is detailed in Chapter 6.

Chapter 3

Studying Human Query-Biased Summarisation

The human summarisation process consists of "identifying the most important ideas and condensing them in a coherent text" [Irwin and Doyle, 1992]. This process involves other cognitive tasks such as comprehension [Thiede and Anderson, 2003], memory [Garner, 1982], learning, [Kintsch, 1990], and reading patterns [Hyönä and Nurminen, 2006], to mention a few. Therefore, these summaries can be considered as general abstracts. Brown and Day [1983] defined a framework that describes three broad processes or "macrorules" that people follow to create summaries: *selection, condensation* and *transformation.* After reading a text, a person evaluates the source content to select the parts for inclusion in a summary, while others are ignored. Relevant material identified in the selection process is not usually used verbatim, since people condense it with general ideas or with more specific concepts. Subjects then transform the remaining abstracted ideas by integrating and combining them. Condensation and transformation are the product of intellectual tasks, which are complex to emulate through an algorithm. Thus, automatic summarisation methods aim to replicate the selection process.

In this chapter we explore how summaries towards a specific information request are created by humans, that is, query-biased summaries. This is our first research question. We conduct a small-scale user study, where participants are asked to produce two query-biased summaries: the first is a generative query-biased summary, as it demands participants to create new prose and may involve any cognitive process above described; and the second is an extractive query-biased summary, where participants are asked to select any part of a

document to create a summary. This study aims to compare human query-biased summaries against automatic approaches, and to find deficiencies in such methods. Automatic extractive summarisation methods studied in this chapter consists of ranking sentences in documents, where the highly scored sentences are taken to form query-biased summaries. We propose to evaluate these sentence ranking methods by introducing eye tracking data collected from participants while creating their generative query-biased summaries, and by using the typical bag-of-words overlap. We detail particular objectives of this chapter in the next section.

3.1 Human Query-Biased Summarisation

We investigate a generative and extractive approach of human query-biased summaries. In short, we refer to them in this chapter simply as summaries. Given an information request and a document, participants are asked to create two short summaries towards the request: a generative summary and a extractive summary. For a *generative* summary, participants can write anything they consider important. These summaries are an expression of a participant's preferred or ideal query-biased summary of a particular document and information request. In contrast, for an *extractive* summary participants take verbatim parts from documents such as sentences or phrases. That is, participants create their query-biased summaries by concatenating parts of the document rather than attempting to produce new prose.

Typical tasks carried out by participants in psychological experiments include composing short general summaries in response to instructions such as *"write a summary of the text"* [Brown and Day, 1983; Garner, 1982; Winograd, 1984]. Evaluation of automatic summarisation methods relies on asking assessors to select sentences in a document that are representative of a document [Edmundson, 1969; Goldstein et al., 1999; Jing et al., 1998; Marcu, 1997; Rath et al., 1961], or on writing general natural language summaries as in DUC/TAC tasks. However, we note that little research has investigated human behaviour involved in producing query-biased summaries. For example, former research requested people to select sentences that they deem to be important for a set of Web pages according to an information request [Varlamis and Stamou, 2009; Wang et al., 2007]. Others have employed summaries of newspaper articles to derive in a set of patterns of generic summaries [Goldstein et al., 1999]. However, patterns obtained from generic human summaries may not be applicable to inform query-biased summarisation methods. Given the limited previous work that examines the construction of query-biased summaries from a subject's point of view, we explore the following hypotheses regarding our first research question.

- We explore whether humans create extractive summaries from the same text fragments that they read when building their generative summaries. We hypothesise that an extractive summary should have a high resemblance with its generative counterpart. Due to subjects combining or generalising ideas to produce a synthesised thought, it is difficult to locate in the document the fragments that are used to construct a summary. Jing and McKeown [1999; 2000] proposed a Hidden Markov Model in order to align human summaries with the corresponding text fragment. However, there is still a language barrier in human and automated summaries to detect these fragments given paraphrases, synonyms, and vocabulary in common between relevant and non-relevant parts of a document [Brown and Day, 1983; Hutchins, 1987]. We propose to use eye tracking techniques to monitor participants' reading patterns to determine more accurately text fragments employed in query-biased summaries. Based on eye tracking evidence, we compare human generative and extractive summaries. We call this approach a position-dependent analysis. Results regarding this hypothesis are detailed in Section 3.4.1.

- Following the position-dependent analysis, we examine whether current algorithms for ranking sentences towards query-biased summaries select ideal fragments referred in participants' generative summaries. In particular, we study the following approaches to automatically generating summaries: a cluster of significant words method; a query term occurrences method; and a combination method that employs the former two methods and introduces a position bias score. Section 3.4.2 describes results of comparing automatic methods against human query-biased summaries.

- We evaluate the performance of automatic methods by ignoring eye tracking data, that is, a position-independent analysis. This analysis (also called a bag-of-words analysis) assumes that a good automatic query-biased summary contains the same words as in a human query-bised summary, regardless of where in the document these words occur. Position-independent approaches are typical ways to gauge summary effectiveness, as described in Section 2.8.1. We investigate if the performance of automatic methods was overestimated when measuring vocabulary overlap, compared to a position-dependent evaluation approach. This is detailed in Section 3.4.3.

We continue to describe our experimental setting for the user study.

3.2 Experimental Design

Summarisation is a cognitive process being studied in several areas of Psychology. We first examine factors that can impact people while constructing generic summaries, in order to minimise these effects in our study. For example, we attempt to minimise participants' fatigue when carrying out the experiment tasks or to engage participants more easily.

Former research has identified three main factors influencing the human summarisation process [Hidi and Anderson, 1986]: text features, functionality and conditions. People can struggle with certain text features when writing a summary, such as document length, genre (narratives or expository text), elaborate sentence structure, complex vocabulary, or lack of structure. On the other hand, the functionality of a summary depends on the target audience rather than the document itself. A writer-based summary serves its author for own purposes, while a reader-based summary benefits a wider audience, is limited in length, and is written after careful inspection of a document. The last factor relates to the conditions under which a summary is constructed. For example, whether people create summaries with either the presence or the absence of a text, enabling a range of cognitive skills to be studied such as memory [Garner, 1982], comprehension [Thiede and Anderson, 2003] or learning [Kintsch, 1990]. This section details: documents and information requests; participants, ethics and general procedures; and study tasks.

3.2.1 Documents and Information Requests

We used two texts from the *LA Times* subset of the TREC newswire collection, which had two associated information requests (or queries). That is, these documents were returned by a modern search engine for the queries, and judged as relevant for both requests by TREC assessors. Documents did not exceed 1000 words and shared similar features outlined in Table 3.1. To avoid biasing the selection of documents to our perception of complex vocabulary, we employed readability scores to guide the selection of documents, and to determine how close they were to plain English. In other words, how easy the documents were to understand. We computed three readability scores using the Unix command `style`: the Flesch index, the Kincaid index and the Fog index. The Flesch index operates in a scale from 0 to 100, where the higher the score, the easier a document is to comprehend as it is closer to plain English. Other scores provide a scale to classify documents based on grade level as the Kincaid index, or years of education as the Fog index does. According to these scores, the chosen documents have medium difficulty, and show to be understandable by subjects in 9th

Features	Document A	Document B
Paragraphs	6	6
Sentences	30	23
Number of words	555	453
Average sentence length (words)	18.9	18.6
Short sentences ($\leq$ 14 words)	13	11
Long sentences ($\geq$ 29 words)	3	2
Flesch Index (0-100)	66.0	56.7
Kincaid Index (grade level)	8.7	10
Fog Index (years of education)	11.5	13.4

Table 3.1: Details of documents employed in the study.

and 10th grade; undergraduates are unlikely to have problems with these documents.

Participants were asked to create short query-biased summaries according to specific information requests. The general task scenario and the administration of information requests were situated in a hypothetical but feasible context to engage participants in the experimental task [Borlund, 2000]. This approach consists of framing a situation (a general request) in which subjects are involved, and then a task subtly turns to an indicative request that defines the particular participants' enterprise. We employed the title of information requests to assemble simulated work tasks (or general requests). For example, for the topics *Endangered species (mammals)* and *Wildlife extinction*, we created the following simulated work task:

> *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 you to help summarise a document according to specific questions that could arise from the audience.*

We employed the narrative of topics to define indicative requests. These requests were slightly paraphrased according to the document content to accommodate them in the task. We provide an example of indicative requests in our study in Table 3.2, according to the topics above mentioned. In Appendix C, we present the second simulated work task and its set of indicative requests.

3.2.2 Participants, Ethics and General Procedures

We recruited ten volunteers from RMIT University. We followed RMIT University ethics protocols and procedures to recruit participants, and to collect data. Volunteers read a Plain Language Statement to learn about their overall task and their rights as participants. In

Original Request	Simulated Scenario
Description of Topic 304: Compile a list of mammals that are considered to be endangered, identify their habitat and, if possible, specify what threatens them.	Indicative Request 1. 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.*
Description of Topic 347: The spotted owl episode in America highlighted U.S. efforts to prevent the extinction of wildlife species. What is not well known is the effort of other countries to prevent the demise of species native to their countries. What other countries have begun efforts to prevent such declines?	Indicative Request 2. The document should be useful to prepare for the questions: *What efforts have been made to prevent the demise of pandas? Which countries are making such efforts?*

Table 3.2: Indicative requests based on two TREC queries. Since the chosen document includes information about a specific mammal (panda), we modified the indicative requests according to the document content.

addition, they signed a consent form indicating their agreement with procedures of the study and storage of data (product of their participation).

Consistent general instructions were provided to participants to maintain uniformity during the experiment. These included: to keep their head still in front of the eye tracker device; not to use any mobile device while doing the experiment tasks; and not to search or to browse. Participants did a training exercise before carrying out the experiment, where the interface features, task and the eye tracking calibration process were explained. Participants were told to that they should complete the tasks on their own, and that they would receive assistance only in case of a technical disruption with the interface or with the eye tracking device. Any questions about the interface or task were answered during this training.

Once volunteers finished the training, we administered a survey (the full survey is included in Appendix D) to collect demographic information. Our volunteers aged between 18 and 40. Nine subjects were enrolled in a computer science program, while one was in the business school. Eight volunteers had a bilingual or multilingual background with English as their second language, whereas two were native speakers.

After participants completed the survey, we calibrated the eye tracking device. The calibration process consisted of participants looking at a specific target in different screen locations, in order for the eye tracker to recognise a particular subject's eye movements [Tobii Technology AB, 2008]. Aside from this, volunteers conducted their experiment tasks as they normally interact with any other desktop computer. Furthermore, discrete supervision was conducted in order to detect any unusual event with the interface or the eye tracker. However,

no participants requested support during the study. Participant data was collected during two weeks, from August 25 to September 10 2009.

3.2.3 Study Tasks

We designed a user interface to facilitate the task procedures to volunteers. Participants were asked to read a simulated work task, as shown in Section 3.2.1, and then to read a corresponding document. Once they finished reading, the interface displayed one of the indicative requests for which the summary should be created. First, participants had to write a short summary, no more than 400 characters in length, related with the information request. If the summary exceeded 400 characters in length, volunteers had to reduce the content. We assumed that this length was enough for participants to provide a short query-biased summary. For example, the INEX Snippet Retrieval track evaluated automatic snippets of at most 300 characters [Trappett et al., 2011]. We called this a generative summary, since subjects were encouraged to compose prose. Participants stopped writing when they considered that their generative summary was sufficient to complete the task.

We were not aiming to study memory or learning skills while participants wrote their generative summary. Thus, the document and information requests remained displayed during this part of the study. The interface provided a text area beside the document, so volunteers could type their summaries and still consult the document. In order to encourage volunteers to create their generative summaries, the interface did not allow participants to copy and paste text directly from the document. Figure 3.1 displays three main elements of the generative interface: (1) the document area; (2) the information request area; and (3) the typing area. Once they finished with the generative summaries and clicked on the button `Continue`, the interface showed the same information request to remind participants of the current topic of the summary. Then, the interface for the extractive summarisation task was loaded.

In the extractive task, participants were able to select any part of the document, such as whole sentences, phrases or words, which were copied into a summary area located next to the document. Figure 3.2 illustrates the key elements of the extractive summarisation interface. The first two elements were the same as in the generative task. However, this interface included a summary area to copy selected fragments (element 3), and buttons to assist volunteers during the task (elements 4 and 5). Participants could select any fragment of text that did not exceed a whole paragraph and add this to the summary area by pressing

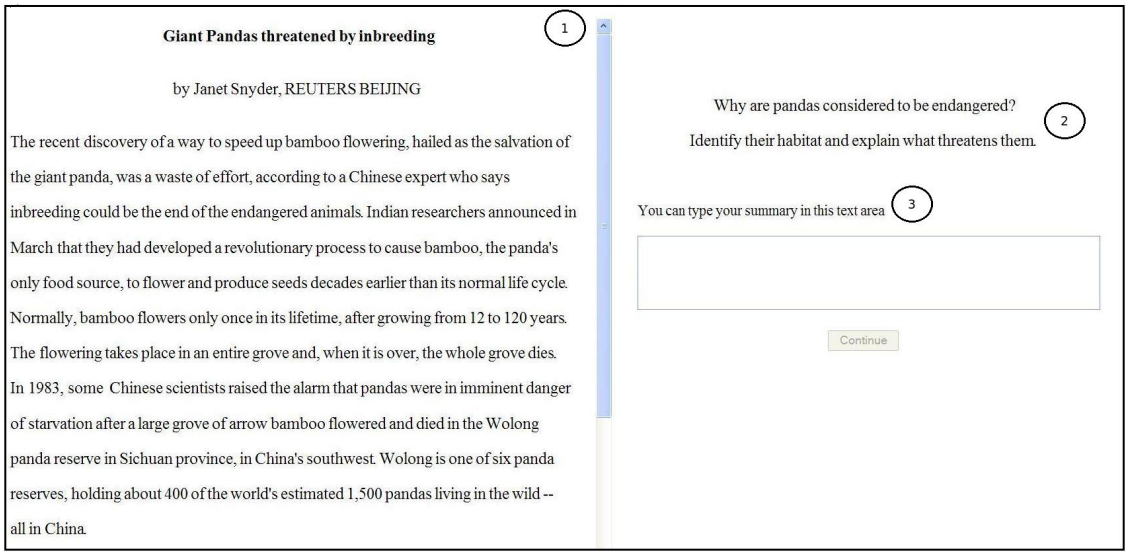

Figure 3.1: Screenshot of the generative interface showing three main elements: (1) the document area; (2) the information request area; and (3) the typing area.

the button "—>>>", or could remove previously selected text by clicking on the button "<<<—". In addition, volunteers could customise their summary by reordering extracted text using the buttons `Up` or `Down`. As can be seen in the figure, once the chosen text fragment was copied into the summary area, it remained highlighted in the document, so participants were easily aware of previous selections. After clicking on the button `Continue`, the interface introduced the next information request, and started the task loop again.

An experiment session involved participants to complete both generative and extractive summaries for two queries and two documents. The order of documents and information requests were randomised and balanced across participants. For each volunteer, we recorded eye movements during the whole experiment session. In general, participants took approximately one hour to complete the experiment. This time included ethics procedures, the training session, the survey, the calibration process, and tasks of the study. Participants were not constrained to finish their generative and extractive summaries in a particular amount of time.

3.3 Tracking Eye Movements

By employing eye tracking techniques, explained in Section 2.10.2, we investigate whether participants construct extractive summaries from the same text fragments that they read

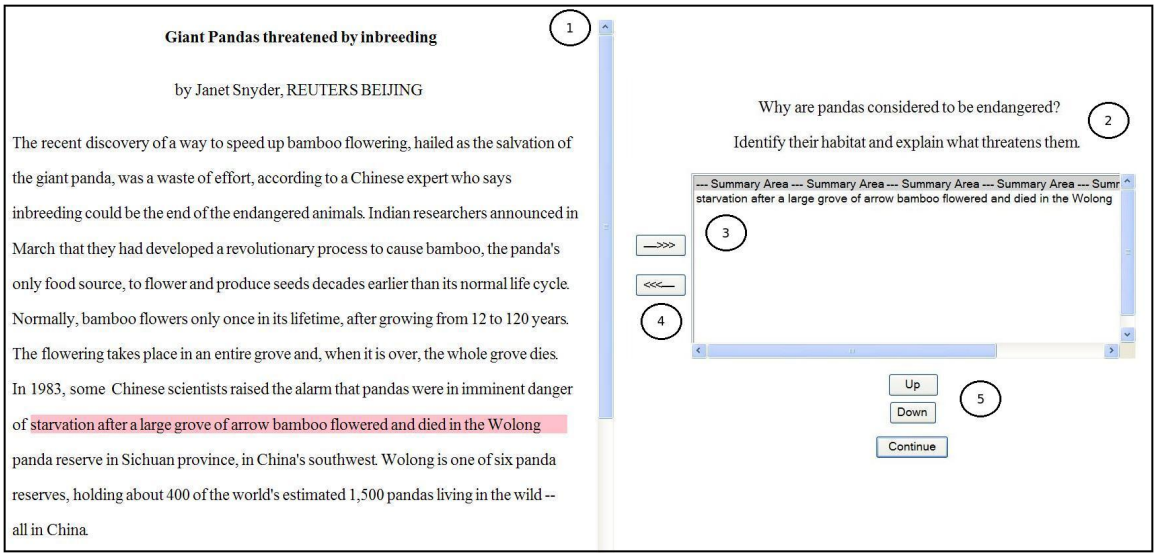

Figure 3.2: Screenshot of the extractive interface displaying five elements: (1) the document area; (2) the information request area; (3) the copy-paste summary area; (4) the buttons to select (—>>>) or to remove (<<<—) text into or from the summary area, respectively; and (5) the buttons `Up` *and* `Down` *to customize the order of chosen fragments.*

when building their generative summaries. In Section 2.8, we described alignment algorithms, which aim to detect original sentences in a text that initiated or conveyed similar meaning with a sentence written by a subject [Jing and McKeown, 1999; Marcu, 1999]. These automatic approaches rely on vocabulary overlap; however, we suggest that these approaches may identify sentences or fragments of text that are not necessarily used to construct a query-biased summary. We continue to describe eye tracking settings, and the procedure we followed to map query-biased summary sentences to text fragments using eye tracking data.

3.3.1 Eye Tracker Device Setting

To record eye movements of participants in our study, we used a Tobii T60 eye tracker.[1] The tracker device is integrated with a 17" TFT display, a resolution of 1280x1024 pixels and a frequency of 60Hz. As mentioned in Section 3.2.2, we had to calibrate the Tobbi T60 prior to volunteers carrying out the experiment tasks.

The Tobii T60 is equipped with a special software, Tobii Studio, which helps in the analysis of eye movements. Depending on the stimuli type such as visual search or reading, Tobii Studio provides filtering mechanisms to identify fixations according to temporal and spacial

[1] `http://www.tobii.com/`

information of each gaze [Tobii Technology AB, 2008]. That is, fixations are characterised by their duration and size. Recall that eye movements consist of fixations and saccades, see Section 2.10.2 in Chapter 2. We analysed our data using Tobii Studio version 1.2 and the ClearView Fixation filter with particular settings for reading. For reading filtering, Tobii Studio recommends to set fixations duration to 40 miliseconds and fixation radius to 20 pixels [Tobii Technology AB, 2008].

3.3.2 Mapping Eye Movements

For each generative summary, we identified the set of text fragments in a document that were used to construct that summary. Reading behaviour is characterised by a contiguous path of forward or backward fixations; backward eye movements are a manifestation of encountering a difficult piece of information [Rayner, 1998]. Longer saccades in the same or different lines generally indicate that participants were scanning the text [Campbell and Maglio, 2001]. For each participant, we analysed recordings and gazeplots, these allowed us to identify potential parts of documents that were used to construct generative summaries. This assessment was carried out in two stages, as we detail below.

First, we identified the fragment of text that was read immediately prior to participants typing in the summary area. We considered a reading area as a sequence of contiguous fixations from left to right in a document line. In particular, fixations located in reading areas had a total fixation duration from 0.20 to 45.20 seconds. Large saccades away from the text region or to the left were allowed in the mapping of reading, as long these saccades corresponded to reaching the end of a line to keep reading further in the text. The author judged the content of the mapped reading area a corresponding or not to the generative summary content.

In some instances it was not obvious which part of the text was read immediately prior to constructing a generative summary. This was due to some participants relied on memory from the initial reading to build their summary rather than reading the displayed document again. Hence, it was not possible to identify the source fragments for some portions of generative summaries. As future work, we recommend to include external judges to improve this assessment exercises. We detail collected eye tracking data in Table 3.5 of Section 3.4.1 for both types of summaries, and for both documents and queries. Figure 3.3 illustrates an example of fixations and saccades in a gazeplot for a read document fragment.

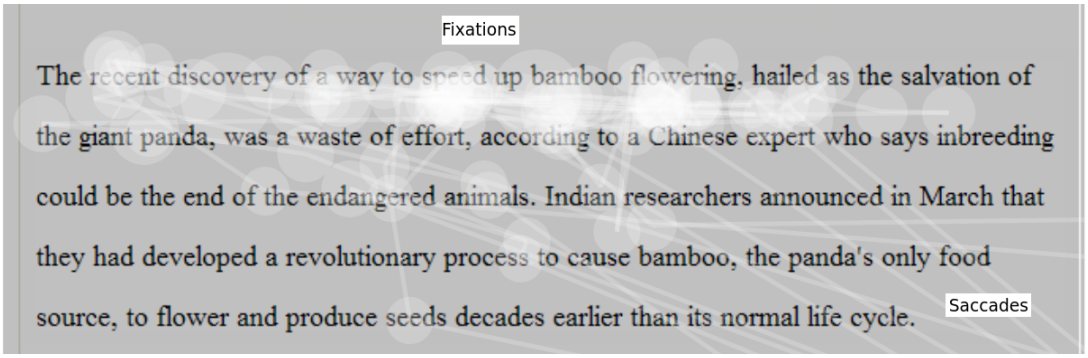

Figure 3.3: Tobii gazeplot for participant 2 while constructing the generative summary for document A and request 2. Overlapping light gray circles represent fixations over the text, and lines denote saccades.

3.4 Analysis of Results

Based on manual assessments of eye movement recordings and gazeplots of participants, we were able to reconstruct generative query-biased summaries based on their corresponding locations in the source documents. This enabled us to measure the overlap between generative, extractive and automatic summarisation approaches. Typically, the evaluation of automatic summarisation techniques consists of measuring vocabulary matches between model summaries (created by humans) and their automated versions. We propose a position-dependent evaluation approach, which relies on eye tracking data of generative summaries. That is, we measured the overlap of fragments of text (word positions) that were used in the generative summary and automatic methods. We compared this with a typical bag-of-words (or position-independent) approach to evaluate automatic summaries.

In our study tasks, as detailed in Section 3.2.3, we did not explicitly requested participants to write complete sentences and such summaries were limited on a character basis during the experiment. Given different writing styles, we found that some participants created generative summaries as a list of "items" per line rather than a cohesive set of sentences. We analysed these items as sentences in order to verify whether the information contained in them was factual. For example, the generative summary created by participant 6, shown in Table 3.4, includes 4 items that do not contain full stops to limit complete sentences. We found that the overall size of human-generated query-biased summaries was 2.98 sentences (or items) with a standard deviation of 0.99. Table 3.3 details the number of sentences for the 10 participants in all document-query combinations.

Participant	A,1	A,2	B,1	B,2
1	2	2	2	2
2	3	3	1	3
3	3	3	2	3
4	4	3	4	5
5	3	2	3	4
6	4	5	3	3
7	3	4	3	5
8	4	3	2	2
9	4	2	1	3
10	3	4	2	2

Table 3.3: Number of sentences (or items) in generative summaries for both documents and queries.

Participant	Generative Summary
6	Starving because bamboo groves die after flowering Man is overcrowding their habitat Inbreeding because there are a few panda males in breeding age Sichuan is Panda's main sanctuary but is also China's most populous province
9	This article talks about several threats to the panda's survival. Pan Wenshi claims that the main threat to their survival is inbreeding. Other threats are mentioned such as poaching and pollution. Pan maintains that the food source (bamboo) is not a major problem in terms of the survival of the panda.
10	There has been an attempt by Indian researchers to enable the Panda only food source of bamboo to be able to flower decades earlier. However others such as Pan Wenshi of Beijing University believe that they had never been in a position to risk starvation. Indian and China have made efforts however China believe that man is a larger risk due to crowding out of the population.

Table 3.4: Example of human-produced query-biased summaries for the indicative request 1: "Why are pandas considered to be endangered? Identify their habitat and explain what threatens them".

By inspecting the content of generative summaries, we noted that four sentences, in the total of 119 written by all participants, were clearly factually incorrect. That is, they were not supported by information in the document. Thus, we removed these sentences (or items) from our analysis. We also corrected misspelled words in generative summaries, since volunteers were not aided with any spelling corrector or electronic dictionary during

the study. In this way, we attempted to reduce any term mismatches that were not likely to occur due to simple misspelling. As an example, Table 3.4 shows three human-produced query-biased summaries for the indicative request 1 detailed in Table 3.2.

3.4.1 Position-Dependent Analysis of Generative and Extractive Summaries

A position-dependent analysis relies on the fact that each term in a text has a unique position, which makes a term different from others. For a position-dependent analysis, we proposed to use the ordinal position of terms within a document to detect those parts that were used in a summary. For instance, the term *panda* located in sentence 3 is different from the term *panda* in sentence 7. A term not only has a different placement in the text, but may also have a different context that surrounds it. Using the eye tracking data, we could distinguish whether participants would construct generative summaries from the same text fragments that they employed to construct their extractive summary — our first hypothesis regarding our first research question. We proceed to explain results based on a position-dependent analysis.

For each participant p, we mapped all term positions in a document d that were read and occurred in a generative summary, in response to a query q. That is, $g_p^{d,q}$ was generated based on eye tracking data as described in Section 3.3.2. For some participants, we could not track the content of the generative summary in the original document. Missing generative summaries corresponded to the following participants in particular document-query combinations: $g_5^{A,1}$, $g_3^{A,2}$, $g_4^{A,2}$ and $g_6^{B,2}$. These participants directly typed their generative summary into the response text area, and did not need to read the source document again.

We defined the set of all positions of the terms used in extractive summaries as $e_p^{d,q}$. Note that we know $e_p^{d,q}$ precisely, since participants directly selected parts from the document d. Table 3.5 shows the size of the position set for both generative and extractive summaries for each participant.

Our first hypothesis tests whether $e_p^{d,q} = g_p^{d,q}$, assuming that $g_p^{d,q}$ is an ideal summary for document d and query q given eye reading patterns of participant p. In order to quantify the performance between different types of summaries, we computed the coverage, which is the proportion that a generative summary is contained in an extractive summary. The coverage, given a position-dependent approach, is defined as:

$$coverage = \frac{|e_p^{d,q} \cap g_p^{d,q}|}{|g_p^{d,q}|} \tag{3.1}$$

Document-Request	A,1		A,2		B,1		B,2	
Participant p	$\|g_p\|$	$\|e_p\|$	$\|g_p\|$	$\|e_p\|$	$\|g_p\|$	$\|e_p\|$	$\|g_p\|$	$\|e_p\|$
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

Table 3.5: Number of term positions that could be identified in the original document for each generative summary (g_p), and the number of term positions used in extractive summaries (e_p). The symbol "–" denotes that term positions were not located for that participant in a particular document-query combination.

The coverage was computed for each participant for a specific document-query combinations. Figure 3.4 presents a boxplot for each document-query pair, showing the values of Equation 3.1 for ten participants. Boxes show the interquartile range; the solid line and the dot in the middle of the boxes represent the median and the mean, respectively; whiskers and open circles are extreme values in the data. As can be seen in the figure, the proportion of the source for the generative summary that was used for extractive summaries is on average above 60% for both documents and queries. Regarding the first hypothesis, we found that the overall mean coverage of text from a generative summary being covered by its corresponding extractive summary is 73%. Thus, extractive summaries of participants cover a high proportion of an ideal generative summary.

It can be the case that $e_p^{d,q} = g_p^{d,q}$ because participant p used an extractive technique to construct a generative summary. That is, they chose not to use more complex sentence structure or different vocabulary that drew together various parts of the text, and simply typed these parts as summaries. In such cases the value of

$$\frac{|e_p^{d,q} \cap g_p^{d,q}|}{|e_p^{d,q}|} \tag{3.2}$$

should also yield values 100%. Figure 3.5 shows that generally this is not the case. While extractive summaries include the underlying text used in generative summaries, the reverse is

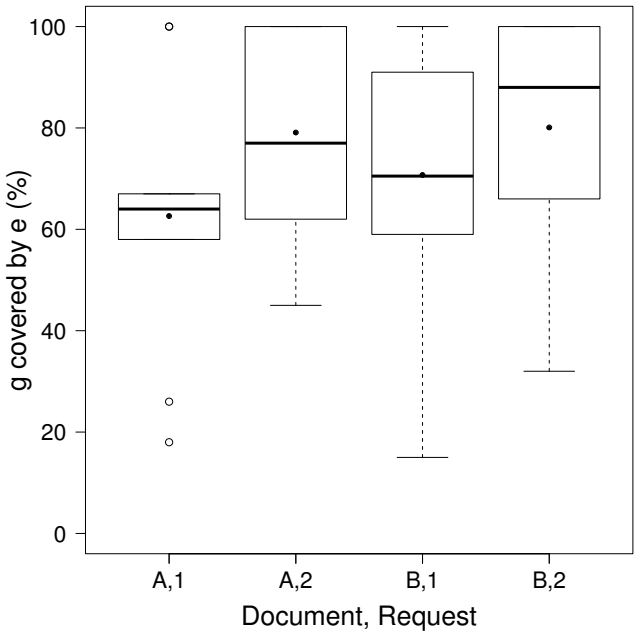

Figure 3.4: Coverage of generative summaries over extractive summaries.

not true. Seven participants used an extractive approach to create their generative summaries for some document-requests pairs (as shown by extreme values close to 100%).

3.4.2 Position-Dependent Analysis of Generative and Automatic Summaries

We continue using the positional-dependent approach to investigate whether automatic summarisation methods select similar information found in participants generative summaries, our second research aim of this chapter. We first describe the automatic methods and then explain results.

Summarisation Methods

We implemented three automatic approaches commonly used in literature to rank sentences. Summaries were comprised of 15% of the documents size, that is, five sentences for Document A and three sentences for Document B. This summary length was proposed by Brandow et al. [1995] and used by Tombros and Sanderson [1998] as a suitable size for summaries. For all methods, we selected the top five and three ranked sentences to construct summaries for Document A and Document B, respectively.

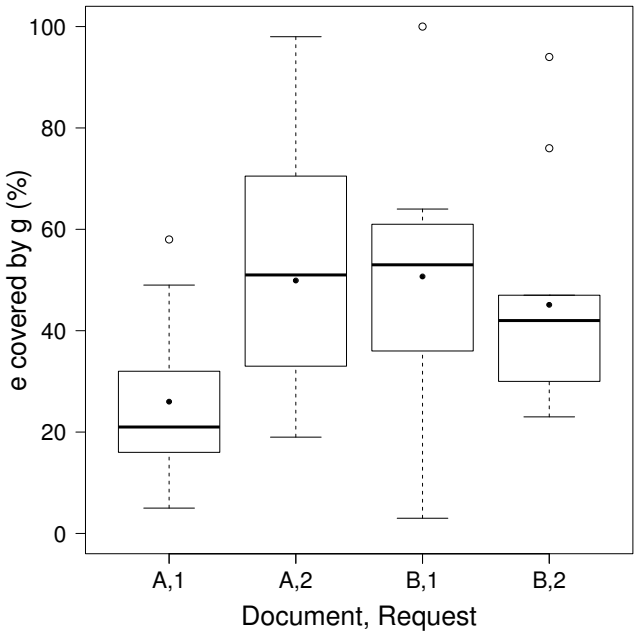

Figure 3.5: Coverage of extractive summaries over generative summaries.

The first method, *CL*, scored sentences according to their clusters of significant words and ignores query terms [Luhn, 1958]. To compile the list of significant terms, we discarded stop words and terms with a frequency less than three [Vanderwende et al., 2007]. The sentence score using *CL* was computed as defined in Equation 2.8 of Section 2.3.1.

The second method, *QB*, included a query term bias [Tombros and Sanderson, 1998]. We submitted the italicised terms of the requests, shown in Table 3.2, as queries. Stopwords were removed from these queries. We used Equation 2.12 of Section 2.4.2 to score sentences.

The third method, *COM*, added a constant *POS* bias [Turpin et al., 2007] to the *CL* and *QB* scores to get a total score for each sentence. The values for the *POS* approach were defined in Equation 2.10 of Section 2.3.1.

Generative vs Automatic Summaries

Figure 3.6 shows the word positions that were frequently selected in extractive summaries by the 10 participants (first panel), those that were read prior to creating generative summaries by the 10 participants (second panel), and those that were ranked by the three automatic summarisers (third panel). The frequencies of selection were normalised in each panel by the corresponding maximum frequency reported in each type of summary. Thus, a progressively

intense blue color indicates that a term was frequently chosen to be included in the summary. Rectangles and triangles in the figure represent words and query terms, respectively. Spaces between a block of rectangles and triangles denote the start of new paragraph. We can observe that in extractive and generative summaries of participants, document parts from paragraphs two, five and six were particularly popular. However, automated methods did not select document sections from paragraphs five or six.

We computed the coverage defined in Equation 3.1 between summaries using the three automatic methods (*CL*, *QB* and *COM*) and generative summaries created by participants. Specifically, $e_s^{d,q}$ denotes the term positions of the summary that a method s has generated from document d and query q. Then, the coverage of Equation 3.1 is re-defined as:

$$coverage = \frac{|e_s^{d,q} \cap g_p^{d,q}|}{|g_p^{d,q}|} \tag{3.3}$$

Table 3.6 shows the set size of terms in the automatic approaches, that is, the position set e for each document-request combination. The size of the generative summaries was detailed in Table 3.5. The coverage described in Equation 3.3 between automatic approaches and generative summaries for both documents and queries is presented in Figure 3.7. Table 3.7 shows the mean coverage between generative and *CL*, *QB* or *COM* summaries, denoted by a dot within each box of Figure 3.7.

A one-way ANOVA analysis of the summary type (*CL*, *QB*, and *COM*) indicated the presence of statistically significant ($p < 0.001$) differences in coverage for each of the four document-request combinations. A follow-up Tukey Honest Significant Difference test demonstrated that, in each case, the participant-created extractive summaries showed significantly higher coverage than any of the automated approaches ($p < 0.001$), while there was no difference between the three automated approaches ($p > 0.100$). At a macro level, there was no significant difference in coverage between the two requests (t-test, $p > 0.100$), or between the two queries for each document ($p > 0.100$ in both cases). This supports our second hypothesis that current algorithms for the construction of query-biased summaries do not select the same fragments that participants read when constructing their ideal generative summaries. While the overall coverage between extractive and generative summaries of participants is around 73%, it is high compared with automatic methods. The mean coverage for both requests in Document A is 26%, and for Document B the mean coverage is 19%. These results are based on eye tracking evidence. We next compare these outcomes with the typical bag-of-words approach for summary evaluation.

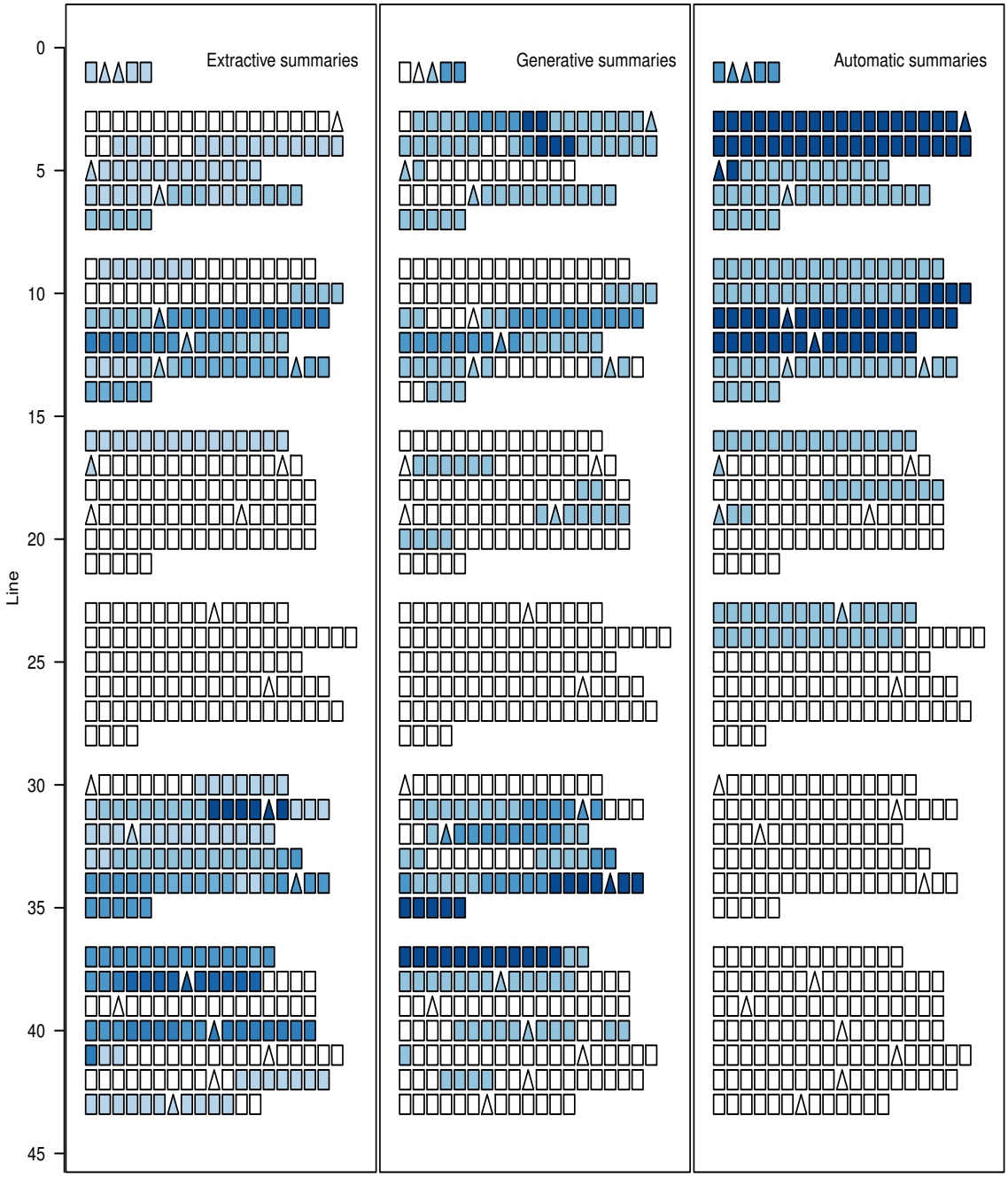

Figure 3.6: Frequency maps of term positions that were used in different summary types for Document A, Request 1. Rectangles denote words in a document, while triangles represent query terms. The Y-axis denotes the number of lines of the document as they were displayed in the eye tracker device. A progressively darker blue color indicates that a word position was frequently selected.

Approach	A,1	A,2	B,1	B,2
CL	138	138	62	62
QB	126	126	60	63
COM	131	131	62	62

Table 3.6: Number of term positions in each automatic summary.

Approach	A,1	A,2	B,1	B,2
CL	27%	8%	12%	3%
QB	32%	17%	1%	43%
COM	37%	36%	8%	43%

Table 3.7: Mean coverage of term positions between generative and automatic summaries.

3.4.3 Position-Independent Analysis

In a position-independent analysis — also called as bag-of-words (BOW) analysis —, a good summary contains the same words as a model summary, regardless of where in the document these words occur. This is the case for automatic evaluation approaches such as ROUGE metrics [Lin, 2004] that measure vocabulary overlap. We conducted a position-independent analysis to gauge the effectiveness between generative, extractive and automatic summaries collected in our user study. We hypothesise that system performance is overestimated when using a position-independent evaluation compared to a position-dependent evaluation.

We defined $G_p^{d,q}$ as the set of words used in a generative summary by participant p for a given document-query pair d, q, defined $E_p^{d,q}$ as the set of words in the corresponding extractive summary, and $E_s^{d,q}$ the set of words in an automatic summary created by method s. A capital letter is used to indicate that the generative and extractive summaries here refer to set of terms rather than positions of terms ($g_p^{d,q}$, $e_p^{d,q}$ and $e_s^{d,q}$) as in previous sections. The coverage between generative and extractive summaries following the BOW approach is given as:

$$\frac{|G_p^{d,q} \cap E_p^{d,q}|}{|G_p^{d,q}|} \tag{3.4}$$

while the coverage between generative and automatic summaries is:

$$\frac{|G_p^{d,q} \cap E_s^{d,q}|}{|G_p^{d,q}|} \tag{3.5}$$

Prior to calculating the coverage, we removed stopwords from extractive, generative and automatic summaries. The first box in each panel of Figure 3.8 shows the coverage as defined

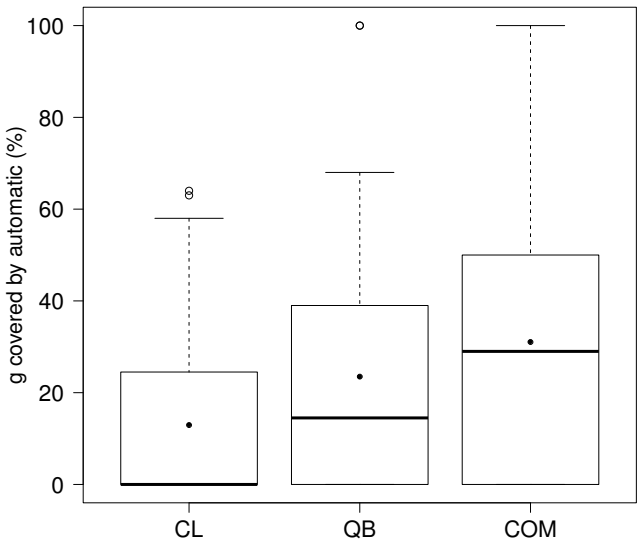

Figure 3.7: Coverage based on term positions of generative summaries against automatic methods for both documents and requests.

Approach	A,1	A,2	B,1	B,2
Extractive	34%	47%	60%	52%
CL	29%	18%	16%	13%
QB	26%	23%	15%	36%
COM	33%	30%	16%	35%

Table 3.8: Mean coverage based on term overlap between generative and automatic summaries.

in Equation 3.4, and the remaining three boxes of each panel represent the coverage as defined in Equation 3.5. Details of the mean coverage is displayed in Table 3.8.

Based on a BOW approach, the automated methods generally performed better, showing a higher coverage, while the participants performed worse. Note that the coverage between generative and extractive summaries created by participants was 48%, and automatic methods slightly improved performance reaching around 24% of the coverage. In particular, for Document A with request 1, shown in the left-most panel of Figure 3.8, there were no significant differences between any summary types (ANOVA, $p > 0.100$). Each of the other three document-topic pairs showed significant effects (ANOVA, $p < 0.001$). For Document A with

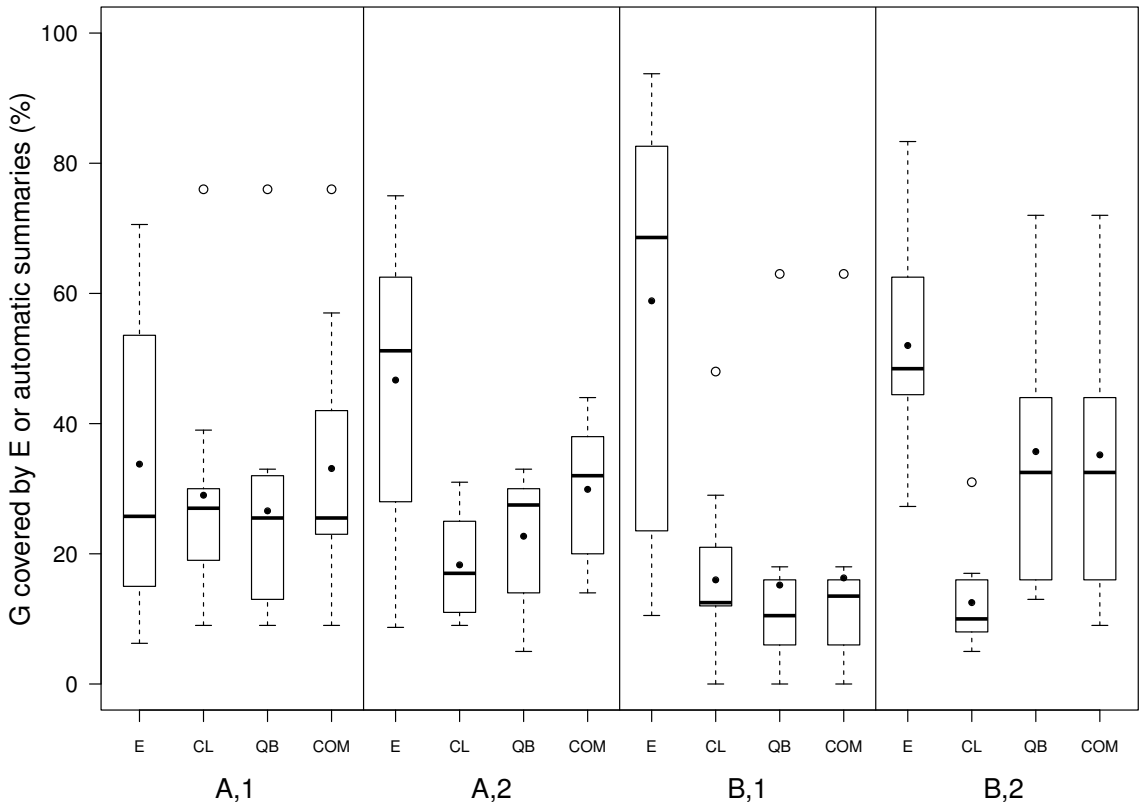

Figure 3.8: Coverage based on bag-of-words approach of generative and extractive summaries (E), and coverage between generative and automatic summaries (CL, QB, COM).

request 2 (second panel), *CL* and *QB* were significantly worse than the participant-created extractive summaries based on the Tukey HSD test ($p < 0.001$ and $p = 0.002$, respectively), while the extractive summary and the *COM* method were close to the traditional significance threshold ($p = 0.051$). For Document B with request 1 (third panel), participants' extractive summaries showed significantly higher coverage than automatic methods ($p < 0.001$), while differences between the automatic methods were not significant. Finally, for Document B with request 2 (fourth panel), the *CL* approach significantly performed worse than all other approaches ($p < 0.001$ for *CL* versus extractive, and $p = 0.022$ and 0.025 when comparing *CL* with *QB* and *COM*, respectively). Again, a *t*-test showed no significant differences in coverage between the two documents overall ($p = 0.966$), or between the two topics in Document A ($p = 0.762$) or Document B ($p = 0.192$). It is clear from the preceding analysis that conclusions about relative levels of coverage between different summary types are affected depending on whether coverage is measured using the eye-tracking data (position-dependent)

or the BOW approach (position-independent).

We observed that a key factor influencing the low coverage in the position-independent approach was the word selection in generative summaries. Recall that subjects may turn to generalise concepts or ideas during the creation of generative summaries, and this can involve using different vocabulary. We examined sentences in generative summaries, and found that 44 out of 119 written sentences contained terms that did not have any correspondence with the document vocabulary. For example, an original sentence of Document A discusses one of the reasons that may affect the shrinkage of the panda population. Three participants included in their generative summaries the fact that the man is approaching the panda's habitat; however, they employed the term *human* instead of *man*, as shown below:

Document sentence	Generative summary sentence
Pandas, which in prehistoric times ranged virtually across China from Beijing to China's extreme south, have shrunk in numbers at the approach of man.	(1) Human development takes place in their own natural environment (2) Human activity affected their environments (3) Threaten by human being

3.5 Discussion

In Section 3.4.1, we explained the relationship between human-created extractive and generative summaries. Recall that not all of the generative text was chosen to be part of the extractive summary. That is, the proportion given by Equation 3.1 was not 100%. There are several possible reasons for this. First, sometimes the extractive summary is shorter than the generative summary, which means the proportion can not be 100%. From Table 3.5 we can see this occurred in 7 instances (for example, $|g_5^{B,1}| > |e_5^{B,1}|$). Second, one of the participants exhibited a possible learning effect, where the first request for each document had a low coverage proportion, but the second was high. As the order of documents and requests was randomized and balanced across all participants, this should not have an overall effect, but it was interesting to observe. Specifically for participant 9:

	$\|g_9\|$	$\|e_9\|$	Equation 3.1	Equation 3.2
Document A, 1st query	57	86	26%	17%
Document A, 2nd query	54	55	100%	98%
Document B, 1st query	27	130	15%	3%
Document B, 2nd query	32	78	100%	41%

Finally it was possible that our extraction of g_p was incomplete due to limitations of the accuracy of the eye-tracking system for a small number of cases. Table 3.9 shows the ratio of the number of words in $g_p^{d,q}$, the generative summary constructed by us using eye tracking data, and $G_p^{d,q}$, the actual summary that a participant typed. This data is graphically presented in Figure 3.9. Generally our identified summaries were about the same size as the typed summaries, with an overall ratio mean of 0.91. This gave us some confidence that we did not introduce length-based artifacts into our analysis. However, it was sometimes difficult to get accurate eye tracking data for participant 3, and that may contribute to low ratios. That is, we were not able to accurately identify all text regions in the source document for $g_3^{d,q}$.

We also computed the ratio by ignoring stopwords in case these terms influence length-based artifacts. We found that this was not the case, as the overall ratio mean was 0.94, similar to that reported above. Details of ratios for different document-query combinations discarding stopwords are presented in Table 3.10 and Figure 3.10.

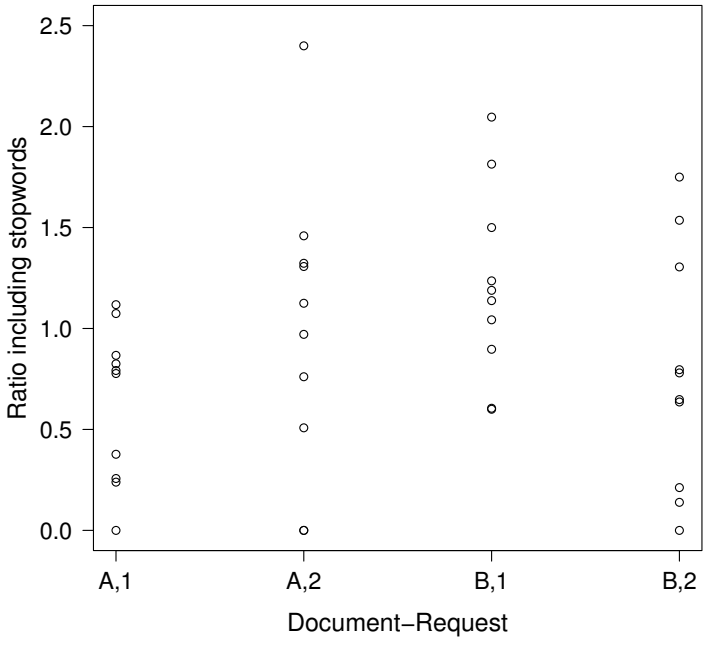

Figure 3.9: Ratio of term positions of generative summaries including stopwords.

p	A,1	A,2	B,1	B,2
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

Table 3.9: The ratio $|g_p^{q,d}|/|G_p^{q,d}|$ for each of the four d, q combinations and 10 participants. This calculation includes stopwords.

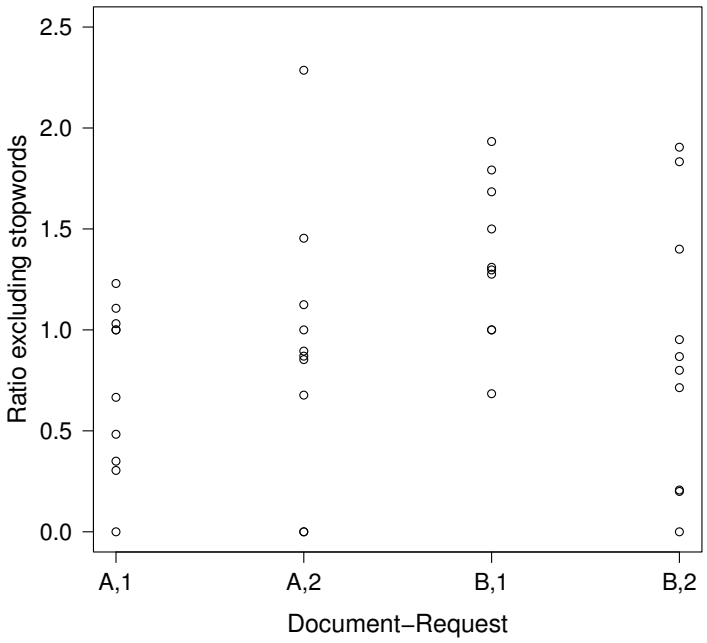

Figure 3.10: Ratio of term positions of generative summaries excluding stopwords.

p	A,1	A,2	B,1	B,2
1	1.000	0.895	0.684	0.714
2	0.666	1.125	1.500	0.800
3	0.304	0.000	1.000	0.200
4	1.000	0.000	1.310	0.868
5	0.000	0.870	1.792	0.207
6	0.350	0.677	1.684	0.000
7	0.483	1.000	1.933	1.833
8	1.230	2.286	1.000	1.400
9	1.107	1.454	1.276	0.952
10	1.031	0.853	1.297	1.905

Table 3.10: The ratio $|g_p^{q,d}|/|G_p^{q,d}|$ for each of the four d, q combinations and 10 participants. This calculation excludes stopwords.

Participants took more time when creating generative summaries (a mean time of 5 and 3.5 minutes when working with documents A and B, respectively), compared to extractive counterparts (3.5 and 2 minutes for documents A and B, respectively). We randomised and balanced documents and queries between participants; however, extractive summaries were always done after generative summaries. Hence, participants were likely able to detect fragments of information faster for the extractive task. While it is cognitively more demanding to produce prose than to select fragments of text, we asked participants to complete the generative summary task first, as a way to avoid biasing subjects to write about what they had previously selected for an extractive task.

Participants in our study were instructed to choose any piece of text that did not exceed a paragraph. We observed that participants carefully narrowed the information in their extractive summaries. That is, less than 50% of the selections made were whole sentences, as subjects commonly ignored prepositions, articles, adverbs and common verb conjugations. Another common pattern we found of subjects is that they assessed documents from top to bottom, and they selected pieces of text to construct extractive summaries following that pattern. Summarisation methods may benefit from preserving the order of extracted sentences to match with the order in the original document, even when a sentence score is higher or lower with respect to their ordinal positions. We construct query-biased summaries for a user study, detailed in Chapter 6, following this observation.

3.6 Summary

In this chapter we presented results from an exploratory user study that investigated human behaviour while constructing generative and extractive query-biased summaries. These summaries were then compared against automatic approaches to analyse how humans and algorithms select sections of the same document. Generative summaries encouraged subjects to produce new prose, while extractive summaries were constructed from excerpted parts of a given document. By using eye tracking techniques, we found that the parts of a document that were used to create extractive summaries were the same as those parts that participants read when constructing generative summaries around 73% of the time. However, automatic summary approaches did not tend to select document parts from the same areas employed by participants. Our findings pointed out that generative summaries were covered by automatic methods around 22%. In fact, the criteria that an extractive summary should be drawn from the same term positions in the document as a generative summary is much more restrictive than typical summary evaluation measures. Automatic sentence ranking approaches could not capture this content by relying on short users requests, as these did not have occurrence in the original document. In the next chapter, we study query expansion techniques to minimise the vocabulary gap between requests and document content to rank sentences to assemble a query-biased summary.

Evaluation of summaries usually relies on measuring overlapping vocabulary. Following a bag-of-words evaluation approach, we discovered that human extractive summaries performed worse, while some of the automated methods performed more strongly. In particular, the coverage of generative and extractive summaries was 48%, and generative against automatic methods was 24%. It therefore appears that relying on eye tracking data, the position-independent analysis overestimate performance of automatic summaries compared against generative summaries.

A small number of documents, requests, and the use of eye tracking techniques can be possible limitations to formally inform query-biased summarisation approaches. In the next chapter, we address this shortcoming by using a large scale testbed to evaluate sentence ranking methods.

Chapter 4

Improving Query-Biased Summaries with Query Expansion

Extractive summarisation techniques commonly rely on excerpting passages of documents that are assumed to be important according to certain attributes. Typically, passages take the form of sentences, since they can convey single and complete ideas. Thus, a key component of extractive summarisation techniques is sentence ranking, as the top m sentences are selected, concatenated and presented as a summary. Results of our previous pilot experiment pointed out a vocabulary gap between the content of query-biased summaries produced by humans and by extractive state-of-the-art methods. This vocabulary mismatch problem is well-known in document retrieval, where IR systems have employed query expansion techniques to help reduce this gap.

In this chapter we investigate the generation of query-biased summaries as a sentence ranking approach, and adopt document ranking techniques to assemble and to evaluate these summaries. This aim is related to our second research question. First, we investigate the effectiveness of statistical query expansion to improve the selection of prospective relevant sentences. While several query expansion approaches have been explored extensively for document retrieval [Billerbeck, 2005; Buckley et al., 1995; Carpineto and Romano, 2012; Efthimiadis, 1996; Voorhees, 1994], few efforts have been focused on summarisation techniques. For example, Sanderson [1998] found that query expansion is not useful for selecting passages where the information request is fully detailed, whereas Losada [2010] discovered that expansion approaches are effective for passage retrieval task involving sentence ranking methods given a set of documents. This chapter examines several sentence ranking techniques

to generate short query-biased summaries, which rely on succinct information descriptions to provide an indicative summary. In addition, we study query expansion in a single-document sentence ranking context rather than in a group of documents.

Second, by using the Novelty track data (described in Section 2.9.1) we adopt the Cranfield methodology, which is widely employed to gauge the effectiveness of document retrieval systems [Cleverdon, 1967], to evaluate the studied sentence ranking methods. This testbed is appealing to our purposes, as it contains assessments of sentences that are relevant according to a specific topic. Therefore, relevance assessments can be used to replicate the Cranfield methodology for evaluating sentence ranking methods. This chapter describes our experimental setting, and discusses results of sentence ranking and query expansion approaches using the Novelty track dataset.

4.1 Sentence Ranking Assisted with Query Expansion

Query-biased summarisation techniques generally favour the selection of passages or sentences that contain query terms. These type of summaries are common in search engines. Typically, search engines display a *caption* [Clarke et al., 2007] for each result returned. This caption includes a title of a document, a short query-biased summary — called a *snippet* — and a URL pointing to the location of a document. The snippet provides users with clues that may characterise a document as a potential relevant result, which is worth further examination. In fact, previous research has found that query-biased summaries can guide users more accurately to find relevant documents [Tombros and Sanderson, 1998; White et al., 2003]. In commercial search engines, research has focused on the appearance and features of captions and snippets that trigger users to click on certain results [Clarke et al., 2007; Cutrell and Guan, 2007; Kaisser et al., 2008; Kanungo and Orr, 2009; Rose et al., 2007]. In other types of applications such as archives [Fachry et al., 2010], bibliographic catalogs [Marcus et al., 1978], and early hypertext browsing tools [Egan et al., 1989], a query-biased summary is a core component that users employ to guide their searches without fully reading all documents they encounter. Thus, we investigate approaches to improve the sentence selection to make up short query-biased summaries. The aims of this chapter are three-fold.

- We study four approaches for ranking sentences: a clustering of significant terms [Luhn, 1958]; a linear combination; a query term occurrence heuristic [Tombros and Sanderson, 1998]; and a vector space model approach applied to the context of sentences [Allan et al., 2003]. The first method ignores users information requests, while the last three

approaches are query-dependent methods. The four methods explore different sources of evidence to rank sentences. These methods establish not only a baseline for further comparison, but also a guide to discern useful features to rank sentences towards the construction of query-biased summaries.

- Similar to document retrieval, we employ query expansion approaches to boost the identification of relevant sentences. Specifically, we use statistical expansion techniques based on relevance feedback and pseudo-relevance feedback (see Sections 2.6.1 and 2.6.2 for a description). In summarisation techniques or passage retrieval assisted by query expansion, the look-up of extra terms can be performed over the entire content of assumed relevant documents [Amini et al., 2005; Han et al., 2000; Losada, 2010; Sanderson, 1998], or can be reduced to a fine-grained set such as sentences [Goldstein et al., 1999; Ko et al., 2008; Losada, 2010]. Our main goals are to find whether the expansion is useful for single-document query-biased summarisation techniques, that is, if sentence ranking methods can be benefited through an expanded query, and which expansion techniques are effective towards a query-biased summarisation task.

- Similar to previous work [Losada, 2010; Metzler and Kanungo, 2008], we use the Novelty track data as our testbed. By following traditional methodologies in document retrieval, such as the Cranfield methodology, we aim to gauge sentence ranking assisted by query expansion.

4.2 Experimental Setting

The Novelty track is a compilation of topics, documents, and relevance assessments at the sentence level. That is, sentences in documents were judged as relevant or non-relevant. Section 2.9.1 fully describes the composition of the Novelty track dataset from 2002 to 2004. Recall that organisers advised to consider data of the Novelty track 2002 as an exploratory study. Consequently, we used in our experiments data from the next two years. This section explains how we used topic information to create baseline and expanded queries. We also describe particulars of the four sentence ranking approaches studied in this thesis.

4.2.1 Baseline Queries

Assessors of the Novelty track created topics that consist of a title, a description and a narrative. Topic titles for both Novelty tracks 2003 and 2004 average three words, similar

Field	Before pre-processing	After pre-processing
Title	Microsoft antitrust charges	microsoft antitrust charg
Description	What are the opinions on Microsoft's guilt or innocence on charges of antitrust?	microsoft guilt innoc charg antitrust
Narrative	To be relevant, the topic needed to be about guilt, not just remedies or settlements. If both guilt and settlement were mentioned it was relevant. Opinions of the general public expressed through opinion polls and interviews, through editorial opinions, and through lawyers for the Department of justice were relevant. Opinions from Microsoft legal experts as reported in the press, opinions from witnesses for both sides and opinions expressed by the judge trying the case were relevant.	topic need guilt remedi settlement gener public express poll interview editori lawyer depart justic microsoft legal expert press wit side judg try case

Table 4.1: Title, description and narrative fields of the topic 14 in the Novelty track 2003 before and after pre-processing.

to current Web queries length [Bendersky and Croft, 2009]. We employed terms in the title field as our main baseline query for sentence ranking methods. The title, identified as t in our experiments, can serve to mimic real users requests. The concatenation of the title with either the description ($t+d$) or the narrative ($t+n$) forms two more baselines, which resemble users being more verbose in the specification of their information needs.

The title, description and narrative fields received a pre-processing, which included case-folding and the removal of stopwords (as explained in Section 2.1.1). Further, common instruction words (i.e. *find*, *describe*, *explain*), and sentences of the narrative field that specified information not required by the topic were deleted. The remaining terms were stemmed, also described in Section 2.1.1, using the Porter algorithm [Porter, 1980]. An example of a topic before and after pre-processing is displayed in Table 4.1.

4.2.2 Expanded Queries

To investigate the impact of query expansion on sentence ranking methods, we explored a relevance feedback technique proposed by Rocchio [1971], and a pseudo-relevance feedback technique, LCA, proposed by Xu and Croft [1996; 2000]. Sections 2.6.1 and 2.6.5 described these approaches in detail. In the literature we identify that the query expansion

process for summarisation or passage retrieval can be applied at two levels: at the *document level* [Losada, 2010; Sanderson, 1998], where an initial ranking of documents is required to gather extra terms; and at the *sentence level* [Ko et al., 2008; Losada, 2010], where an additional sentence ranking step is carried out using top ranked documents. In the following sections, we explain the document level and the sentence level approaches using either Rocchio's technique or LCA.

Expanded Queries at the Document Level

Relevance feedback methods may require an initial subset of relevant and irrelevant documents to create an improved version of the original query. This information can be usually collected from users interacting with IR systems [Rocchio, 1971]. However, in the absence of users providing such assessments, an initial ranking of documents can be employed instead, where the top documents are assumed to be relevant. Local analysis techniques then rely on mining extra terms from this set of documents.

The Novelty track dataset, examined in Section 2.9.1, lacks information about the rank position of each document, since assessors could conduct multiple searches to identify useful documents. In order to obtain an initial ranking of documents, we indexed the AQUAINT collection using the open source search engine *Zettair*.[1] We employed the Okapi BM25 ranking function, and submitted the pre-processed title field to generate an initial ranking of documents for each topic. We selected the top five documents pertaining to the Novelty track that occurred in the ranking provided by the search engine, and defined these top documents as the relevant set, R'. These documents were then pre-processed in the same way as the topic statements. We used the content of documents in the relevant set R' to expand the query. Hence, we call this expansion process at the document level. We proceed to describe implementation particularities of Rocchio and LCA approaches at the document level.

Rocchio's Approach at the Document Level. The first method that we employ for query expansion is the relevance feedback method proposed by Rocchio [1971]. This approach formalises queries and documents as vectors, while relevance information is used to return a new re-weighted query vector as shown in Equation 2.14 (page 27). In this formulation, an initial query Q_0 is submitted to an information retrieval system. Based on the returned results list, information about the relevance or non-relevance of particular documents is obtained,

[1] `http://www.seg.rmit.edu.au/zettair/`

either through explicit feedback provided by users, or by making a pseudo-relevance feedback assumption. We followed the second approach, since it did not rely on user assessments.

The formulation allows the setting of three parameters to control: the relative influence of the original query terms, α; terms from relevant documents, β; and terms from non-relevant documents, γ. Literature suggests the weights of these parameters as α=8, β=16 and γ=4 [Croft et al., 2009]. Since we followed a pseudo-relevance feedback assumption, we proposed that the original query terms were preserved in full (overriding α) given our aim to generate summaries towards users requests. Specific identification of non-relevant documents in the Novelty track was not available; thus, the parameter γ was set to zero. In our approach the parameter β then can be any constant value, as it did not affect the final term weighting. Therefore, in our implementation of Rocchio's approach at the document level, expansion terms were based on their relative frequency of occurrence in the five top-ranked documents R'. We called these *Rocchio-D terms*, as the whole content of documents in R' was used to expand the query. Section 4.4.2 presents results of *QB* and *VSM* methods using this expansion technique.

Local Context Analysis at the Document Level. Local Context Analysis aims to weight words in the top ranked documents, according to their co-occurrence with query terms [Xu and Croft, 1996; 2000]. To apply LCA, we then re-used the set R' — also known as the local set — to obtain the expansion terms. Xu and Croft [1996] proposed an expansion based on concepts, which can be represented as noun-phrases or single words. The LCA technique weights terms in a entire document or passages in a document to gather expansion words. Losada [2010] explored LCA at the document level for passage retrieval tasks and proposed two modifications: concepts are single terms, and sentences are passages. We implemented LCA following these modifications to rank sentences given a document rather than passage retrieval tasks as Losada [2010] investigated. The function that weights a term t in a given query Q $(q_1, \ldots, q_m)$, and its corresponding components are re-defined [Losada, 2010]. This was done in order to manage the "term" notation; instead of concepts. Equations 2.16, 2.17 and 2.18 in page 33 are re-written as:

$$co\_degree(t, q_i) = \log_{10}(co(t, q_i) + 1) * \frac{idf(t)}{\log_{10}(n)} \tag{4.1}$$

$$co(t, q_i) = \sum_{s \in S} f(t, s_j) \cdot f(q_i, s_j) \tag{4.2}$$

The $co(t, q_i)$ component in Equation 2.16 involves the number of co-occurrences of the term t and q_i in sentences of the top documents $S = \{s_1, \ldots, s_n\}$. These are given by $f(t, s_j)$ and $f(q_i, s_j)$, respectively.

In our experiments $idf(t)$ represents the inverse frequency of term t in a collection of N sentences. That is, the total number of sentences either in the Novelty track 2003 or 2004. Specifically, from topics 1-50 N=39,820 and from topics 51-100 N=52,447. These values were reported in row four of Table 2.5. The parameter N_t is the number of sentences in the collection where the term t appears. The parameter n corresponds to the number of sentences contained in each local set R', in other words, the sentences in the first five documents for a given topic.

$$idf(t) = min(1.0, \log_{10}(N/N_t)/5.0) \tag{4.3}$$

Recall that LCA weights concepts as shown in Equation 2.15. Thus, the function depicting the co-occurrence degree of terms is given below:

$$f(t, Q) = \prod_{q_i \in Q} (\delta + co\_degree(t, q_i))^{idf(q_i)} \tag{4.4}$$

Terms weighted according to Equation 4.4 are named *LCA-D terms*. The letter D stands for the document level expansion. Section 4.4.2 shows results based on the expansion at the document level using *LCA-D* terms.

Expanded Queries at the Sentence Level

Rocchio-based and LCA-based query expansion techniques can also be carried out at the sentence level. That is, after retrieving the top documents, the individual sentences in these documents are ranked. Consequently, this approach weights terms using the top sentences and ignores other sentences in the document. Previous research has pointed out that this approach is suitable for the construction of search engine snippets [Ko et al., 2008] and passage retrieval [Losada, 2010]. We investigated whether expansion using only the top documents is as effective for locating additional terms as including an extra sentence ranking step.

Having identified the top five documents as explained in previous experiments, the Okapi BM25 ranking function was used to rank sentences towards the topic title, and the first five sentences in each document were employed in the expansion process. Minimal modifications were required to apply Rocchio or LCA query expansion techniques at the sentence level. For Rocchio expansion, the component $\vec{d}$ in Equation 2.14 is substituted by $\vec{s}$, which represents

a vector for a given sentence. For the LCA approach, now the collection consists of the top five documents, and the local set only considers the first five ranked sentences. We name these approaches as *Rocchio-S* and *LCA-S*, respectively, since the local set R' for locating extra terms is reduced to sentences. The performance of sentence ranking methods using this expansion scheme is discussed in Section 4.4.3.

4.2.3 Summarisation Methods

We studied four summarisation methods: a clustered approach (*CL*); a query-biased approach (*QB*); a linear combination (*COM*); and a ranking document technique applied to the sentence context (*VSM*). The first method is query-independent, whereas the other three are biased towards the query. The first and third method were described in Section 2.3.1, while the other two in Section 2.4.1. In this section, we briefly outline these methods, and provide details of our implementation.

The *CL* approach. To assemble summaries for the *CL* approach, we followed Luhn's method to score sentences according to clusters formed by significant and non-significant terms, as detailed in Equation 2.8 (page 16). We obtained significant words for each document by discarding stopwords, as well as terms with a frequency below three, as suggested by Vanderwende et al. [2007]. If clusters were not formed for a given sentence, the sentence received a score of zero.

The *QB* approach. The *QB* method is based on the occurrence of query terms in sentences. The score of a sentence is defined by Equation 2.12 (page 23). The computation of the *QB* score did not involve repetition of query terms in a sentence.

The *COM* approach. This approach employed three sources of evidence to rank sentences: a cluster component, a query-biased component and a position component. The three components were weighted using a linear combination, with the total score of a sentence being calculated as:

$$COM_s = w_1 \cdot CL'_s + w_2 \cdot QB_s + w_3 \cdot POS_s \tag{4.5}$$

where $w_1 = w_2 = w_3$ are weights controlling the relative contribution of each component. We explain in Section 4.4.1 the values assigned to each weight.

The CL' component was a modified cluster score based on the term significance selection given by Equation 2.9 (page 17). The QB component tailored the selection of sentences to the information needs of users, and is shown in Equation 2.12. The POS component aimed to detect good candidate sentences for summaries based on the ordinal position of sentences in a document. Leading sentences have been shown to be more beneficial for summaries than sentences that occur later in a document [Brandow et al., 1995; Radev et al., 2004]. We used Turpin et al. [2007] weight definition of early sentences as defined in Equation 2.10 (page 20).

The *VSM* approach. In order to study another query-biased ranking approach, we employed the Vector Space Model adapted for sentence retrieval as shown in Equation 2.11 (page 23). In our experiments, we assumed that a query term was not repeated in an information request. That is, Equation 2.11 was simplified by ignoring the component $\log(f_{t,q}+1)$, as we discarded duplicate query terms. In a single-document summarisation task, as we aim, the parameter n in the equation below is the number of sentences in a document, instead of the number of sentences in a collection as employed by Allan et al. [2003]. Thus, the similarity between a sentence and a query was reduced to:

$$R(s|q) = \sum_{t \in q} \log(f_{t,s} + 1) \log\left(\frac{n+1}{0.5 + sf_t}\right) \tag{4.6}$$

Any of the four methods return the top m ranked sentences as summary. In the case where sentences scored the same value, the method resolved ties by using the ordinal position of a sentence within a document. Thus, it gave preference to sentences close to the beginning of a document. Finally, if all of the sentences in a document scored zero when applying either CL, COM, QB or VSM, the m leading sentences of that document were presented as the summary instead. We explain the setting of the m value in the next section.

4.3 Evaluation

The Cranfield methodology [Cleverdon, 1967] is the traditional approach to evaluate effectiveness of IR systems, as it sets a framework containing topics, documents and relevance assessments. Relevance assessments are human judgements given to documents considered useful to one or multiple topics. Thus, these assessments are used in conjunction with particular measures to quantitatively gauge retrieval techniques. In particular, the Cranfield

methodology involved four multi-level relevance judgements identified as: whether the document completely answers a request, highly relevant documents, useful documents containing supporting information of the request, and documents of minor interest.

The Novelty tracks 2003 and 2004 supply relevance judgements at the sentence level, as explained in Section 2.9.1. However, these judgements were not classified in different utility relevance levels. Rather, a sentence is a assessed as relevant or irrelevant. Therefore, similar to document ranking evaluation, we use a Cranfield style methodology on binary judgements to assess the performance of sentence ranking methods. In comparison to typical summary evaluation methodologies, the Cranfield methodology can be considered an intrinsic approach to evaluate summaries, as discussed in Section 2.8.1. That is, the set of relevant sentences can be seen as potential components of a human summary. While there are multiple metrics to gauge document ranking [Croft et al., 2009], these may not be entirely suitable to a sentence ranking approach. We employed P@m as our evaluation measure, where m indicates the sentence cut-off value of returned top ranked sentences. The P@m measure (explained in Section 2.8.1) computes the proportion of m sentences returned by the system, which were also judged as relevant by a Novelty track assessor.

Previous research in extractive summarisation recommends excerpts to contain up to 15% of the document sentences [Brandow et al., 1995; Tombros and Sanderson, 1998]. Both the Novelty 2003 and 2004 testbeds have a mean document length of around 30 sentences; therefore, query-biased summaries can be comprised of 4 sentences. However, we argue that the amount of information conveyed in 4 sentences may not be appropriate for the generation of query-biased summaries due the amount of information being presented. We aimed to generate short query-biased summaries, so we restricted the selection of the two top scored sentences, that is, m=2. We varied this restriction further in Chapters 5 and 6.

A topic in the Novelty track includes multiple documents. For this reason, we averaged the P@2 for each document in a given topic. Finally, we obtained the means for all topics to reduce it to a simple value, and to make comparisons between different sentence ranking approaches. Thus, results reported in this thesis represent the macro averaged P@2 over topics.

We noted Novelty track assessors could not identify relevant sentences for some documents. A possible explanation is that a document may not contain relevant information relative to the topic, or that a document may have overlapping content with another previously assessed document. In addition, irrelevant documents were introduced in the Novelty track 2004; however, sentences in such documents were not assessed. Prior to computing P@2,

Features	Novelty tracks		`sub2003`	`sub2004`
	2003	2004		
Number of topics	50	50	50	50
Number of documents	1,250	1,808	1,120	1,070
Number of sentences	39,820	52,447	35,966	33,456
Number of relevant sentences	15,557	8,343	15,490	8,193

Table 4.2: Statistics of documents in the Novelty 2003 and 2004 tracks, and their subsets `sub2003` *and* `sub2004` *. These subsets satisfy the condition of containing at least two relevant and two non-relevant sentences in every document.*

we chose documents having at least two relevant and two non-relevant sentences. Therefore, documents outside this range may obtain P@2 values that would lead to a possible bias in the averaged measure. For example, evaluating P@2 for a given document only containing one relevant sentence would make hard to compare the performance, since the maximum achievable P@2 score for that document would be 0.5. Table 4.2 shows the statistics of the subset of documents for both Novelty tracks 2003 and 2004. These subsets were used in our evaluation, and for labelling purposes, we call them `sub2003` and `sub2004`. The next section focuses on describing results of sentence ranking methods, both non-assisted and assisted by query expansion.

4.4 Results

We investigated four summarisation methods as detailed in Section 4.2.3. We created a simple baseline using the *CL* approach, analysed the importance of components of the *COM* approach and compared them against exclusive query-biased approaches *QB* and *VSM*. We explored sentence ranking methods that could benefit from employing query expansion techniques. Using the Novelty track dataset, we present results of experiments conducted with these four sentence ranking methods without applying any expansion technique. This baseline enabled to evaluate whether sentence ranking approaches benefit after applying different query expansion techniques. The following sections analyse and discuss our findings.

4.4.1 Sentence Ranking Methods without Query Expansion

The *CL* approach is query-independent and is used as a lower bound on performance for all approaches. The P@2 values for the *CL* method are shown in the first row of Table 4.3. From the four studied methods, the *COM* approach seems the most appealing due to in-

cluding three sources of evidence that are commonly used to weight sentences for extractive summaries. That is, clusters of significant words (CL'), query terms (QB) and position of sentences (POS). Previous research has weighted these components equally [Radev and Fan, 2000; Tombros and Sanderson, 1998; Turpin et al., 2007], that is $w_1 = w_2 = w_3 = 1$ for Equation 4.5 in Section 4.2.3. In initial experiments, we observed that this simple combination approach performs poorly; in fact, it performed worse than using only the query-biased component ($w_1 = w_3 = 0$ and $w_2 = 1$). Intuitively, it makes sense that several sources of evidence should contribute differently to the final score. For example, the positional score gives more priority to the leading sentences in a document. Thus, the *COM* algorithm does not retrieve sentences from other sections of the document if the weight w_3 is too high.

Using the `sub2003` data as a training set, we explored the parameter space for the weights, optimizing them in the range $0 \leq w_i \leq 1$ using increments of 0.025. The optimal weights found were $w_1 = 0.050$, $w_2 = 1$ and $w_3 = 0.025$. We employed this values when using documents of the `sub2004` dataset. Table 4.3 shows P@2 values for the `sub2003` and `sub2004` data using the *COM* method with uniform and optimised weights. In the same table we report P@2 scores for the *QB* and *VSM* approaches. The outcomes reported in the table correspond to employing the topic title as a query for the *COM*, *QB* and *VSM* methods. Optimised weights in the *COM* approach led to significantly better results than uniform weights (paired Wilcoxon test, $p < 0.001$). The percentage increase through using optimised weights was 10% and 12% for `sub2003` and `sub2004`, respectively. Therefore, we continue our analysis employing the optimised *COM* version.

The *COM* method using optimal weights significantly improved over the *QB* method (paired Wilcoxon test, $p < 0.001$) for the Novelty track 2003 dataset. Moreover, the *VSM* approach slightly improved over the optimised *COM* method, the differences were not significant ($p = 0.186$ for the `sub2003`, and $p = 0.955$ for the `sub2004`). While the *COM* method provided evidence of being a potentially effective approach, the query-biased component strongly dominated the other two: clusters of significant words and sentence position. That is, $w_1 = 0.050$ and $w_3 = 0.025$ were very small in comparison to $w_2 = 1$, which corresponded to the query-biased component. Therefore, in subsequent analysis we did not study in the optimised *COM* approach for the following reasons:

- The enhancement of the *COM* method against *QB* was not robust, since these results were not statistically significant for the `sub2004` dataset (paired Wilcoxon test, $p = 0.068$).

Method	Query	sub2003	sub2004
CL	—	0.48	0.34
COM Uniform weights	*t*	0.60	0.47
COM Optimal weights	*t*	0.66	0.52
QB	*t*	0.61	0.52
VSM	*t*	0.68	0.53

Table 4.3: Results of averaged P@2 over topics of the CL, COM, QB and VSM methods. The t character denotes the use of the topic title as a query.

- A possible shortcoming of the *COM* approach is the requirement to pre-calculate optimal weights for each of the components involved. This parameterised process may be sensitive to different collections.

Given the above explanation of the optimised *COM* method, we investigated the role of query expansion in sentence ranking methods such as *QB* and *VSM*. We noted that using the sub2003 data and the baseline title as query, *VSM* significantly outperformed *QB* (paired Wilcoxon test, $p < 0.001$). We continue to explain the effects of query expansion in the *QB* and *VSM* methods in the next section.

4.4.2 Query Expansion at the Document Level for Sentence Ranking

In Section 2.6.5 we describe two query expansion approaches that can be used to improve sentence ranking techniques: at the document level and at the sentence level. We employed these two approaches to expand the query, in particular through that proposed by Rocchio [1971] and Xu and Croft [1996; 2000]. Extra terms were sourced from the top five ranked documents returned in response to the initial query, that is, expansion at the document level. In this section we examine results of the sentence ranking methods *QB* and *VSM* assisted by these query expansion techniques (*Rocchio-D* and *LCA-D* terms).

We compared summaries employing query expansion against summaries produced with the title of a topic (t), the title and the description (t+d), and the title and the narrative (t+n). These query baselines were two-fold: the baseline title served to measure the gain after using query expansion; and the other two helped to compare a manual expansion against an automatic approach. We call these "manual" expansions to the t+d and t+n baselines, as the description and the narrative field were prepared by a Novelty track assessor. In fact, the performance of sentence ranking methods using the t+n baseline can be seen as target to which automatic expansion approaches would aspire.

Query expansion approaches can be parameterised to find an optimal number of documents and terms required in the expansion [Billerbeck and Zobel, 2003; Losada, 2010]. We did not study the optimal amount of documents; we only investigated performance of sentence ranking methods when the title is progressively concatenated with extra terms. That is, adding a different number of expansion terms to the title of a topic with either *Rocchio-D* or *LCA-D* terms, with a step size of five terms. Thus, the title baseline was expanded with batches of five terms at a time until reaching a maximum expansion of 100 terms.

The Figure 4.1 shows the performance of the three baseline queries and expansion techniques for both ranking methods *QB* and *VSM*. Panels (a) and (b) display results for the `sub2003` collection, while panels (c) and (d) do the same for the `sub2004` collection. As can be seen in the figure, P@2 increased as more terms were added into the title. In order to gauge the effect of automatic expansion approaches, we selected 45 terms as a cut-off value to summarise P@2 values for both sentence ranking methods and expansion techniques. The figure shows that after this cut-off parameter the increase in performance among expansion methods is very low. We also noted that in the `sub2004` data the P@2 scores remained flat after 45 extra terms. Moreover, we observed that around 45 extra terms were required among ranking methods and expansion techniques. Thus, this value can be regarded as a robust cut-off to describe our results among collections and sentence ranking methods. Table 4.4 details P@2 performance of both summarisation methods using baseline queries *t, t+d* and *t+n* (rows 1-3), and expanded queries employing 45 extra terms (rows 4 and 5). The first row lists the title baseline results, which have been reported in the previous section; we include them in this table for comparison purposes. Percentage differences between baseline and expanded queries are shown in Table 4.5.

After applying the expansion, overall results suggest that sentence ranking methods improve performance compared to methods that only use the title t as query. Depending on the summarisation-expansion combination, the percentage varies from 10% to 25% for the `sub2003` dataset, and from 10% to 16% for the `sub2004` dataset. While the expansion did not always improve significantly, it at least equaled performance. Summaries of approximately 80% and 60% of the topics were improved when the *QB* and *VSM* methods employed expansion, respectively. Figure 4.2 displays the difference of P@2 values calculated between summarisation methods using the expansion and the title baseline (t). These results demonstrate that query expansion techniques at the document level helped to identify sentences that were judged as relevant by Novelty track assessors. Therefore, we assume that these sentences represent good candidates to create a short query-biased summary of a document.

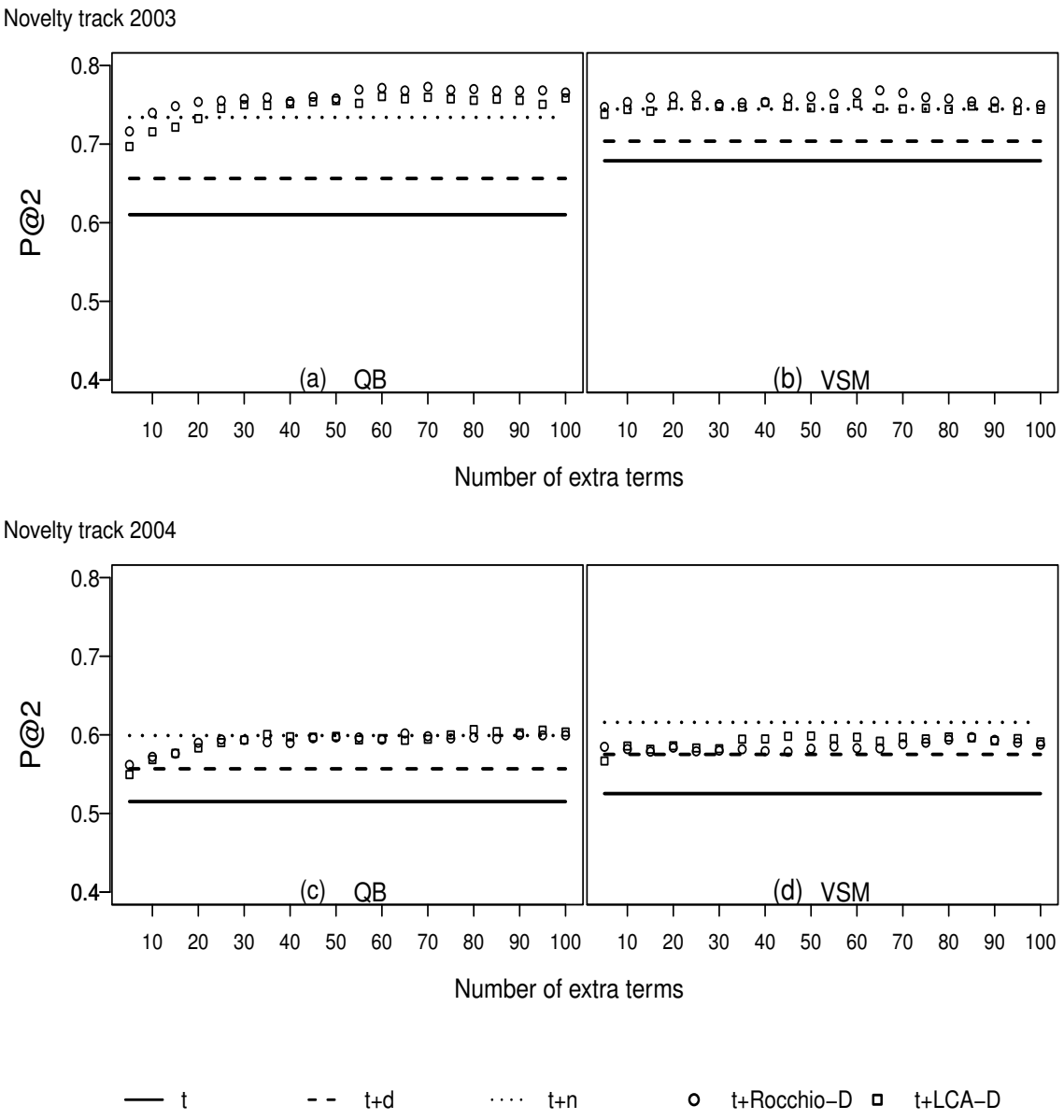

Figure 4.1: Sentence ranking methods assisted by document-based query expansion. The x-axis displays the number of extra terms added to the title, and the y-axis indicates the averaged P@2 scores over topics. The straight line represents the title baseline, while the two dotted lines correspond to the title being concatenated with the description and narrative. Circles and squares denote Rocchio-D and LCA-D query expansion approaches, respectively. Results in panels (a) and (b) used the `sub2003` *dataset, and panels (c) and (d) the* `sub2004` *dataset.*

		QB		*VSM*	
	Query	sub2003	sub2004	sub2003	sub2004
Baseline	t	0.61	0.52	0.68	0.53
	t+d	0.66	0.56	0.70	0.58
	t+n	0.73	0.60	0.74	0.62
Expanded	t+*Rocchio-D*	0.76	0.60	0.76	0.58
	t+*LCA-D*	0.75	0.60	0.75	0.60
Oracle	t+*Rocchio-Best*	0.81	0.66	0.82	0.67
	t+*LCA-Best*	0.81	0.67	0.81	0.67

Table 4.4: Averaged P@2 over topics using QB and VSM. The number of expansion terms in Rocchio-D and LCA-D is fixed to 45, and varies per topic for -Best. Sentence ranking methods employ sentence position to resolve ties.

Sentence Ranking	Expansion vs Baseline	sub2003		sub2004	
		Δ	p	Δ	p
QB	***Rocchio-D***				
	t	24.59%	p<0.001	15.65%	p<0.001
	$t+d$	15.84%	p<0.001	6.98%	p=0.011
	$t+n$	3.58%	p=0.067	−0.57%	p=0.836
	LCA-D				
	t	23.53%	p<0.001	15.91%	p<0.001
	$t+d$	14.85%	p<0.001	7.23%	p=0.008
	$t+n$	2.69%	p=0.025	−0.35%	p=0.673
VSM	***Rocchio-D***				
	t	11.77%	p<0.001	10.14%	p=0.005
	$t+d$	7.81%	p=0.003	0.58%	p=0.936
	$t+n$	1.90%	p=0.213	−6.07%	p=0.030
	LCA-D				
	t	10.23%	p<0.001	13.89%	p<0.001
	$t+d$	6.32%	p=0.020	4.01%	p=0.169
	$t+n$	0.49%	p=0.137	−2.88%	p=0.648

Table 4.5: Percentage differences of averaged P@2 over topics for document-based Rocchio and LCA expansion (using 45 extra terms) versus baseline queries. Significance values are from a paired Wilcoxon test. Sentence ranking methods employ sentence position to resolve ties.

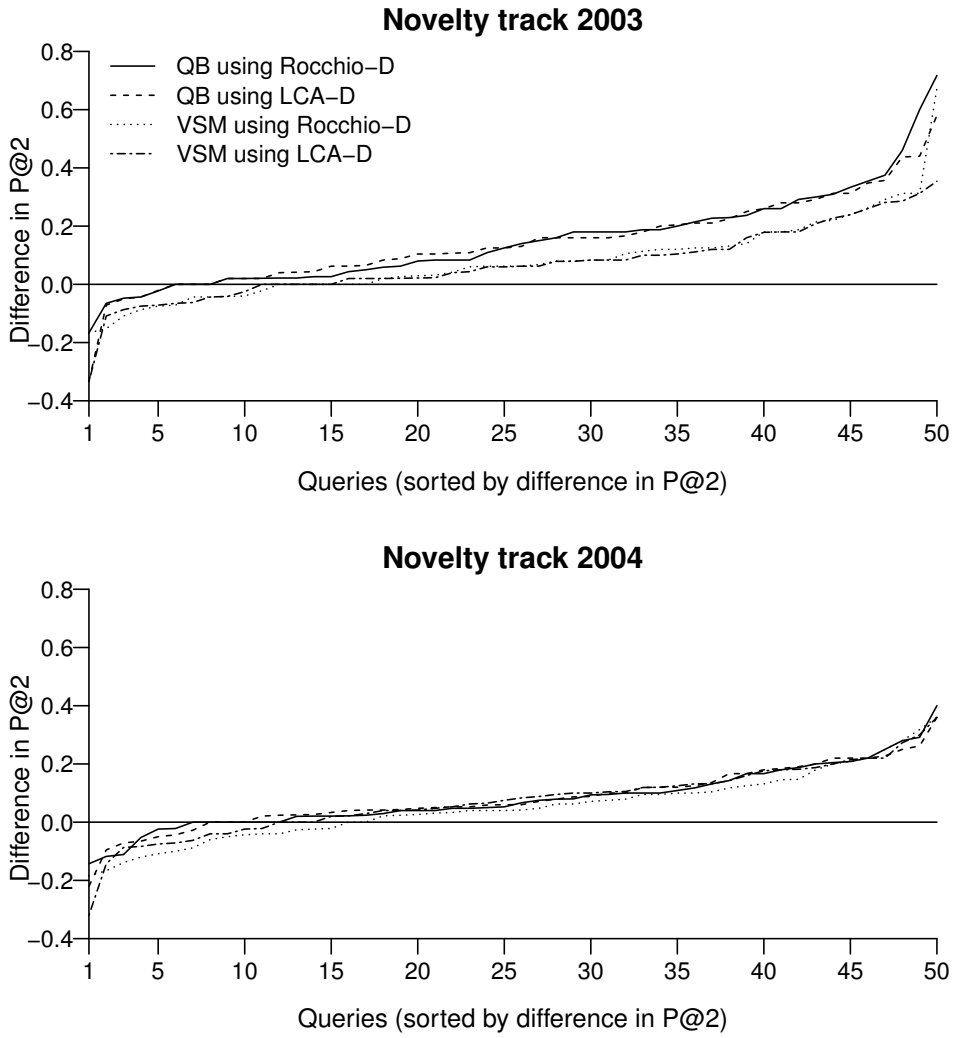

Figure 4.2: Difference of averaged P@2 over topics between sentence ranking methods using the expansion and the baseline title. Note data is discrete, but displayed with lines for clarity of presentation. The graph in the top uses the `sub2003` *dataset, while the graph in the bottom uses the* `sub2004` *dataset.*

In our analysis above, a static collection-wide expansion parameter was used, with 45 terms being added to each topic title. Closer inspection of individual topics showed that this parameter varied on a per-topic basis. For example, to obtain optimal performance employing the *QB* method for the title of topic two (*clone sheep Dolly*) requires 50 extra terms for *LCA-D*, and 25 extra terms for *Rocchio-D*. Based on this consideration, we provide an oracle estimation that indicates the maximum P@2 performance that is obtained through

expansion with an optimal number of extra terms for each topic. These estimation values are identified as *Rocchio-Best* and *LCA-Best* in the last two rows of Table 4.4.

In addition to these results, we analysed the performance of each summarisation-query expansion combination in both datasets. These combinations include: (*QB*, *Rocchio-D*), (*QB*, *LCA-D*), (*VSM*, *Rocchio-D*), and (*VSM*, *LCA-D*). Outcomes revealed that there were no significant differences (paired Wilcoxon test, $p > 0.050$) between any combination using the `sub2003` collection. Nevertheless, for the `sub2004` data there was a slightly significant difference between *QB* and *VSM* with *Rocchio-D* terms expansion (Wilcoxon paired test, $p = 0.034$). This analysis was done using 45 terms. By comparing panels (a) and (b), and (c) and (d) of Figure 4.1, we can appreciate that both summarisation methods behave similarly when using expanded queries. In fact, the performance difference between *Rocchio-D* or *LCA-D* is minimal, or there is no significant difference at all. Observe that this can be applicable to several cut-off values. We noted a difference in the `sub2003` data when using baseline queries, *VSM* outperformed *QB* for the baseline t significantly (paired Wilcoxon, test $p < 0.001$).

We conclude that query expansion at the document level (*Rocchio-D* and *LCA-D*) significantly improves the selection of relevant sentences, and this enhancement is uniform among sentence ranking methods. That is, *QB* and *VSM* report similar P@2 values for both expansion techniques. The next section analyses performance of expansion approaches at the sentence level.

4.4.3 Query Expansion at the Sentence Level for Sentence Ranking

Rocchio-based and LCA-based query expansion can also be carried out at the sentence level, as discussed in Section 2.6.5. This section presents results of experiments using the sentence level expansion to enable a comparison against document level approaches. In order to achieve query expansion at the sentence level, it was required to select the top N documents and then to rank the M sentences within these documents. Extra terms were then selected from these highly scored sentences. Previous research has pointed out that this approach is effective for the construction of search engine snippets [Ko et al., 2008] and passage retrieval tasks [Losada, 2010]. This motivated us to investigate whether expansion using only the top documents can be as effective for locating extra terms, rather than requiring an additional sentence ranking step.

We replicated experiments described in the previous sections. Table 4.6 provides P@2

	QB		*VSM*	
Expansion	sub2003	sub2004	sub2003	sub2004
t+Rocchio-S	0.74	0.58	0.74	0.57
t+LCA-S	0.75	0.59	0.75	0.58
t+Rocchio-S-Best	0.78	0.64	0.79	0.65
t+LCA-S-Best	0.80	0.64	0.79	0.65

Table 4.6: Averaged P@2 over topics using QB and VSM sentence ranking methods. The number of expansion terms in Rocchio-S and LCA-S is fixed to 45. Sentence ranking methods employ sentence position to resolve ties.

scores of sentence ranking methods employing *Rocchio-S* and *LCA-S* (expansion at sentence level) to refine the title topic. We concatenated batches of five terms to the title, similarly to the document-based method. Thus, results reported in this section employed the same cut-off of 45 extra terms for comparison purposes against the document-based approach. However, we noted that sentence-based expansion required slightly less extra terms, around 40 on average for two types of ranking methods and expansion techniques. Moreover, *Rocchio-S-Best* and *LCA-S-Best* showed upper bounds of performance that were achieved using a sentence level expansion.

While the percentage increase was significant between the title and *Rocchio-S* or *LCA-S* expansion approaches, these did not improve against document-based expansion, nor for oracle estimations. Table 4.7 shows the actual percentage change for all baseline queries, and Table 4.8 demonstrates that *Rocchio-S* or *LCA-S* performed more poorly than their *Rocchio-D* and *LCA-D* equivalents although not significantly. The table presents the document and sentence level approaches using the title as the expanded query. Despite of negative changes, these were not statistically significant, excepting for the *VSM* method using *LCA*-based expansion in the `sub2004` dataset. See Table 4.8 to appreciate in detail percentage changes and significance values.

As an example, Table 4.9 presents the first ten terms obtained from *Rocchio-D*, *LCA-D*, *Rocchio-S* and *LCA-S* expansion approaches that were concatenated to the pre-processed topic title 14: "*microsoft antitrust charg*". Note that terms have been stemmed, and those that are different among document-based and sentence-based approaches are in boxes. For this topic, we observed that multiple terms were retrieved by both document and sentence level approaches. For instance, *Rocchio-D* and *Rocchio-S* obtained seven terms in common: `browser case govern justic market monopoli netscap`.[2] At this cut-off value,

[2]Terms are alphabetically listed rather than according to their weights as done in Table 4.9.

Sentence Ranking	Expansion vs Baseline	sub2003		sub2004	
		Δ	p	Δ	p
QB	***Rocchio-S***				
	t	20.45%	p<0.001	12.43%	p<0.001
	$t+d$	11.99%	p<0.001	4.01%	p=0.097
	$t+n$	0.14%	p=0.202	−3.34%	p=0.364
	LCA-S				
	t	22.68%	p<0.001	15.46%	p<0.001
	$t+d$	14.06%	p<0.001	6.81%	p=0.008
	$t+n$	1.98%	p=0.048	−0.74%	p=0.870
VSM	***Rocchio-S***				
	t	9.45%	p<0.001	8.69%	p=0.004
	$t+d$	5.57%	p=0.020	−0.74%	p=0.700
	$t+n$	−0.22%	p=0.227	−7.31%	p=0.035
	LCA-S				
	t	9.87%	p<0.001	10.75%	p<0.001
	$t+d$	5.98%	p=0.003	1.15%	p=0.930
	$t+n$	0.17%	p=0.154	−5.55%	p=0.114

Table 4.7: Percentage differences in P@2 for sentence-based Rocchio and LCA expansion (using 45 extra terms) versus baseline queries. Significance values are from a paired Wilcoxon test. Sentence ranking methods employ sentence position to resolve ties.

10 terms, only 30% of the terms were different. Words such as `1995, expert` and `softwar` can be closely related to the topic or may harm retrieval. This can explain why the sentence level approach did not improve against the document level. We did an analysis to quantify the percentage of similar terms at different cut-off values. That is, after concatenating 10, 20 terms and so on until reaching 100 terms. Overall results showed that *Rocchio-D* and *Rocchio-S* methods selected the same expansion terms approximately 50% of the time, while for *LCA-D* and *LCA-S* approaches the overlap was around 65%. A possible explanation for LCA approaches obtaining a higher number of equal extra terms than Rocchio was that LCA at the document level (*LCA-D*) detected the co-occurrences of query terms with other words within all sentences of the relevant set of documents, R'. Thus, an additional step for ranking sentences (*LCA-S*) produced a set of similar extra terms.

These outcomes demonstrated that including an additional step to rank sentences did

Sentence Ranking	Sentence level	Document level	sub2003		sub2004	
			Δ	p	Δ	p
QB	*Rocchio-S*	*Rocchio-D*	-3.32%	p=0.099	-2.78%	p=0.160
	LCA-S	*LCA-D*	-0.69%	p=0.874	-0.39%	p=0.727
VSM	*Rocchio-S*	*Rocchio-D*	-2.08%	p=0.531	-1.31%	p=0.589
	LCA-S	*LCA-D*	-0.32%	p=0.918	-2.75%	p=0.042

Table 4.8: Percentage change of P@2 when the expansion at the sentence-based against document-based.

Expansion approach	Document level	Sentence level
Rocchio	govern case compani monopoli court market browser jackson netscap justic	govern netscap browser 1995 case expert justic market monopoli softwar
LCA	netscap browser jackson monopoli softwar market govern expert testimoni depart	netscap browser govern corp software monopoli 1995 expert element snare

Table 4.9: Ten extra terms of the topic 14 obtained by Rocchio and LCA using the document and sentence level expansion approaches. Terms in boxes are different from the document or sentence level.

not increase P@2 performance. Using sentences instead of documents may reduce the search space for locating extra terms; however, we argue that the sentence level query expansion approaches (*Rocchio-S* and *LCA-S*) can generate a set of similar terms as document level query expansion methods (*Rocchio-D* and *LCA-D*). We therefore recommend query expansion based on an initial rank of documents, which does not require extra complexity in the expansion process. Based on relevance assessments of the Novelty track, these final selected sentences can potentially be included in a short query-biased summary.

4.5 Discussion

Query expansion has been employed as a way to increase the effectiveness of retrieval systems by refining submitted queries, as users may lack knowledge to formulate an optimal query regarding their information needs. However, drifting the topic of the actual information

request is a common shortcoming of query expansion techniques given the parameterised nature of some approaches [Billerbeck, 2005; Croft et al., 2009; Xu and Croft, 1996]. We discussed in Section 4.4.2 that the number of extra terms to achieve maximum P@2 performance fluctuated among topics. We fixed our experiments to a specific cut-off of extra terms; however, this affected P@2 performance for some topics that required for example 5 extra terms. Thus, adding more expansion terms led to deviate the focus of the original request, and to rank higher non-relevant sentences. We studied the performance of sentence ranking methods when the query is gradually expanded. In some cases the expanded query was longer than sentences in documents. We explore the effects of query expansion related to sentence length in the next chapter. In this work, we have focused on studying the effectiveness of sentence ranking methods involving query expansion. Nevertheless, we did not explore efficiency aspects that query expansion may bring.

DUC/TAC conferences have investigated different summarisation styles and proposed several intrinsic evaluation methodologies since 2001. Such methodologies assessed automatic summaries in terms of vocabulary overlap [Lin, 2004], or content matching units [Nenkova and Passonneau, 2004], against a set of model summaries. These model summaries are usually comprised of abstracts authored by assessors, who may merge ideas into a single sentence or paraphrase content. Hence, the framework provided by DUC/TAC conferences cannot be used straightforwardly to evaluate sentence ranking methods as we aim in this work. We have solely focused on evaluating the sentence ranking methods using the Novelty track dataset. We have gauged effectiveness of those methods based on the sentence topical relevance. Similar to previous research [Metzler and Kanungo, 2008], we also assumed that sentences judged as relevant by Novelty track assessors can be good to assemble query-biased summaries. However, the topical relevance supposition may not entirely fit into the concept of relevant content for constructing query-biased summaries. We investigate this in the next chapter.

4.6 Summary

In this chapter, we studied different sources of evidence to rank sentences and query expansion towards the construction of short query-biased summaries. We evaluated sentence ranking methods by following traditional document ranking methodologies. In particular, we employed relevance sentence assessments of the Novelty track 2003 and 2004 as a realisation of the Cranfield methodology, widely used to assess document retrieval. By studying several

sources of evidence proposed in the literature, we showed that a scoring component that includes query terms is more effective than using other features to rank sentences.

We found that sentence ranking methods were significantly improved by query expansion techniques, and that these ranking approaches performed similarly regardless the type of expansion applied. We discovered that a query expansion approach relying on an initial rank of documents was effective, and did not add complexity. Thus, a more grain-detailed expansion, which required to rank sentences in top documents, did not improve a simple document-based expansion.

Chapter 5

Problems in Summary Evaluation

In the previous chapter we applied a Cranfield-based methodology to evaluate sentence ranking methods using relevance assessments of the Novelty track. Past work has also employed sentence relevance assessments of this track to evaluate passage retrieval [Losada, 2010] and snippet generation [Metzler and Kanungo, 2008]. In this chapter we study two properties of the Novelty track relevance assessments: sentence indicativeness and sentence length. We argue that it is vital to pay attention to these properties of the assessments, in particular when they are used for evaluation in a context that is different from the original aims of the track. We explore how to evaluate sentence ranking methods for query-biased summaries given these two sentence properties, our third research question.

Assessors of the Novelty track provided sentence relevance judgements to study filtering approaches in order to identify relevant information for *ad-hoc* tasks [Harman, 2002; Soboroff and Harman, 2003; Soboroff, 2004], rather than sentence ranking for query-biased summarisation tasks. Thus, these assessments might not be entirely suitable to both generalise and evaluate other applications that are not exclusively concerned with topical relevance. We investigate sentence indicativeness of sentences judged as relevant by Novelty track assessors. That is, although a sentence has been judged as relevant to a Novelty track topic, it may not be a good indicator of the content of the document in a query-biased summary. In other words, the sentence may be relevant, but not indicative. Section 5.1 describes a user study that aims to quantify the proportion of selection where relevant sentences are also indicative, and how to gauge the performance of sentence ranking methods given this proportion of selection.

Examining Novelty track assessments, we observed that sentences judged as relevant tend

to be significantly longer than irrelevant sentences. The sentence length feature was not taken into consideration in generating results in the previous chapter. The length bias in retrieval systems has been studied for documents [Losada et al., 2008; Singhal et al., 1996; Smucker and Allan, 2005] and character-based passages [Callan, 1994], but not for sentences. We study the effects of query expansion techniques to generate short query-biased summaries by introducing a sentence length component in ranking methods. In addition, we propose a novel approach to evaluate sentence ranking methods regardless of the sentence length bias in the collection assessments. We explore the sentence length property in detail in Section 5.2.

In Section 5.3, we investigate whether the sentence length is a factor that can affect the detection of indicative sentences. This observation is based on the length hypothesis studied by Marcus et al. [1978], which states that users tend to select information that progressively increases in size as relevant. We also investigate how both the sentence indicativeness and sentence length affect the perceived performance of sentence ranking methods.

5.1 Sentence Indicativeness

We argue that sentence relevance assessments of the Novelty track are broad in terms of topical relevance, and only a certain proportion cover the requirements to be indicative for query-biased summaries. In this section, we investigate whether a sentence judged relevant by a Novelty track assessor also conveys indicativeness, and how to evaluate sentence ranking methods on an indicativeness basis. *Indicativity* was early described by Marcus et al. [1978] as an attribute to denote the relevance that bibliographic catalog fields convey about a document. As explained in Section 2.7.3, relevance is an abstract concept that generally depends on many factors such as topicality, subjects, task or time, to mention a few. Marcus et al. [1978] questioned the assumption that users solely aim for topical relevance. They stressed that indicativity can be affected by other document features such as readability, novelty, or specificity regarding a topic. Nevertheless, indicativity has not been determined for sentences, and for a particular task such as query-biased summarisation. We define sentence "*indicativeness*", or that a sentence is indicative, as a specification of relevance. In our approach, a sentence is indicative if it is capable of pointing to salient content within a document, and if this content is then useful as part of a short query-biased summary of the document, given an information request.

In order to evaluate automatic summarisation systems through intrinsic methodologies, assessors can be asked to identify a set of sentences within a document that are deemed to

be important or representative for a summary [Edmundson, 1969; Jing et al., 1998; Rath et al., 1961; Sun et al., 2005; Teufel and Moens, 1997; Wang et al., 2007]. Then, automatic methods are compared against this set of representative sentences. However, such efforts are limited in terms of documents, information requests, assessors' expertise, and generally addressed for generic summarisation. In a similar fashion to these intrinsic methodologies, the previous chapter presented results of the effectiveness of sentence ranking methods based on topical relevance given the Novelty track assessments. Past work has also employed relevance sentence assessments of the Novelty track to evaluate snippet generation using machine learning [Metzler and Kanungo, 2008], and passage retrieval assisted by query expansion [Losada, 2010]. However, we suggest that one cannot simply assume that all sentences judged as relevant are helpful for evaluating query-biased summaries. In order to study whether query expansion techniques are useful based on sentence indicativeness, we define two particular aims of this section.

- We conduct a crowdsourcing experiment to study if sentences that were deemed as relevant by assessors also expressed indicativeness to be included in short summaries. Based on the collected data, we aim to quantify the proportion of selection where a relevant sentence is also indicative. We detail the experimental setting and results in Sections 5.1.1 and 5.1.2, respectively.

- We employ the proportion of selection as an assessor error rate (α), which can be used to adjust the effectiveness of sentence ranking methods. Through stochastic simulations using α, we investigate whether sentence ranking methods assisted by query expansion are effective based on sentence indicativeness. We explain our findings in Section 5.1.3.

5.1.1 Experimental Setting

In order to investigate indicativeness of the Novelty track sentence relevance assessments, we required people to judge whether a sentence was a good candidate to assemble a short query-biased summary. To collect indicativeness judgements, subjects were given an information need and a document. Then, participants were shown a pair of sentences from this document, and were requested to select the sentence that they considered best suited for a summary of the present document. This section describes the characteristics of topics, documents and sentences that were used for assessment, as well as the task procedures to collect indicativeness judgements of sentences.

Novelty track	Topic	Description
2003	Topic 2	Cloning of the sheep Dolly
	Topic 4	Egyptian Air Flight 990 disaster in October of 1999.
	Topic 34	John Glenn's Shuttle Discovery trip
2004	Topic 58	Is Irradiated food safe for consumption?
	Topic 74	Find documents that report related information about the car accident that killed Princess Diana on August 31, 1997.
	Topic 82	The first human hand transplant in the United States was performed on Matthew Scott on January 25, 1999.

Table 5.1: Descriptions of six Novelty track topics selected for our user study.

Topics, Documents and Sentences

In Section 2.9.1, we explained the composition of the Novelty track data 2003 and 2004. Given that it was impractical to gather indicativeness judgements for the entire set of document sentences, we chose a certain number of topics, documents and sentences to be assessed. To prevent participant fatigue, we chose topics that had at least five short documents comprising between 15 and 40 sentences. We assumed that this document length reduced scrolling activity, so participants could focus on the judging task instead. From this initial topic filtering, we selected six topics from the Novelty dataset, three for each track 2003 and 2004. We attempted to select topics that may be of interest or easy for subjects to assess, and did not require specialised knowledge, to encourage diligent task completion.

Participants were asked to provide indicativeness judgements towards an information request. Since the topic titles are short, averaging only 3-4 words, we noted that this could increase difficulty of the task to participants. The description field was therefore used instead as information request, and slightly re-phrased to accommodate the task. For example, for the description of topic 74, workers were shown the request as: `Information about the car accident that killed Princess Diana on August 31, 1997`. Table 5.1 lists descriptions of the selected topics, and specific instructions regarding the general task are given further in Table 5.3 in the next section.

Having identified the topics, we manually inspected the content of documents in order to remove those that included repeated information: more specifically, a topic describing the course of an event within a short time frame through different documents. For example, some documents of Topic 4 differed from other documents only by mentioning minor updates about

Document combination	Pair 1	Pair 2	Pair 3
DC_1	(R_1, I_1),	(R_2, I_2),	(R_3, I_3)
DC_2	(R_2, I_3),	(R_3, I_1),	(R_1, I_2)
DC_3	(R_3, I_2),	(R_1, I_3),	(R_2, I_1)

Table 5.2: Three different combinations of sentence pairs given a document, for three relevant and three irrelevant sentences.

the accident such as the status of investigations, reports of victims, or dates, to mention a few. We selected five documents per topic; thus, a total of 30 documents were employed in the study.

In the task, participants were shown a pair of sentences within a document and had to select the best candidate sentence for a short summary. The pair was comprised of one sentence judged as relevant and one judged as irrelevant by the Novelty track assessors. The sentence-paired experiment setting was chosen for three reasons: to reduce the complexity of the task; to encourage attention during the experiment; and to detect indicative sentences regardless of the presence of an irrelevant sentence. From each document selected, we randomly sampled three relevant sentences ($R = \{R_1, R_2, R_3\}$) and three irrelevant sentences ($I = \{I_1, I_2, I_3\}$). This added up to 90 relevant sentences and 90 irrelevant sentences from the 30 documents selected.

We combined each relevant sentence with each irrelevant sentence, and sorted each pair in a way that a relevant sentence was shown at different times during assessments, to avoid any ordering effects. That is, we defined three combinations of sentence pairs, as shown in Table 5.2. For the experiment, 90 different combinations were generated out of 30 documents. We explain the use of these document combinations in the following section.

Task Procedures

As mentioned earlier, indicativeness judgements were collected using crowdsourcing (explained in Section 2.10.3). A Web interface, embedded in the CrowdFlower platform, facilitated the worker's[1] task by displaying the information request and highlighting a pair of sentences at a time within a document, based on a particular document combination setting (DC). Participants clicked on the sentence that they considered to be more indicative of the document. Once a sentence was clicked, the interface automatically showed the next pair. Thus, workers could not change their assessments. For example, a worker who was

[1]We also employ the term *participants* to refer to workers of the crowdsourcing platform.

assigned to assess pairs of DC_2 provided indicative judgements of the pairs: (R_2, I_3),(R_3, I_1) and (R_1, I_2). Recall that pairs were formed by one relevant and one irrelevant sentence. The interface highlighted both sentences with the same color, so participants were not biased in their task. For instance, consider sentences in the pair (R_2=18, I_3=5), since the interface displayed both at the same time within the document, the order between sentences did not affect workers assessments. That is, showing the pair (R_2, I_3) or the pair $(I_3, R_2,)$ had the same display in a document during the experiment. Note that the full document was displayed to provide context when assessing the sentences; however, participants were not required to read the whole text. Table 5.3 shows instructions provided to participants prior to starting the task.

Following the crowdsourcing practices, a unit of work was completed after a participant assessed the three pairs of sentences for a document combination. The number of units in our study corresponded to the 90 document combinations created, as explained in the previous section. Nine document combinations compiled the set of gold units (explained in Section 2.10.3). Recall that CrowdFlower suggests to use between 5% and 10% of additional units being gold, and considers trusted workers to be those who achieve more than 70% of gold units correctly answered.

In our study, a working session was comprised of five units from different document combinations, and participants were able to accomplish a maximum of four working sessions, or 20 units in total. Figure 5.1 illustrates the elements of a unit and a working session in our experiment. We established this limit of working sessions as the task was relatively easy to complete and to mechanise. Thus, the probability of responding correctly to gold questions by chance was greater, so we attempted to reduce (or to avoid) such an effect by limiting the units a worker could complete. Participants assessed one gold unit per working session and, if they missed more than one gold unit they were tagged by the platform as untrusted workers. CrowdFlower does not provide assessments coming from untrusted workers, so we did not include such judgements in our analysis.

Below is a request for information and five documents that contain relevant information for that request. Suppose that you are searching for specific information related to the request. Repeat the following steps for each of the five documents to finish your task.

1. Click on the button "Start".
2. Within the document, two sentences are highlighted at a time. Given the information request, your task is to CLICK on the sentence you would most like to see in a short summary of the document, that would help you decide whether to read the full document or not.

   That is, choose the best sentence that should appear in a summary.

   You will be shown three pairs of sentences, in turn, for each document. Highlighted pairs will be displayed automatically in the documents, and information requests may not be the same among documents.
3. Once you finish selecting sentences in a document, click on the button "Exit".

Note that a successful work task requires the completion of this process for FIVE documents.

Table 5.3: Instructions provided to workers prior assessing indicativeness of sentences.

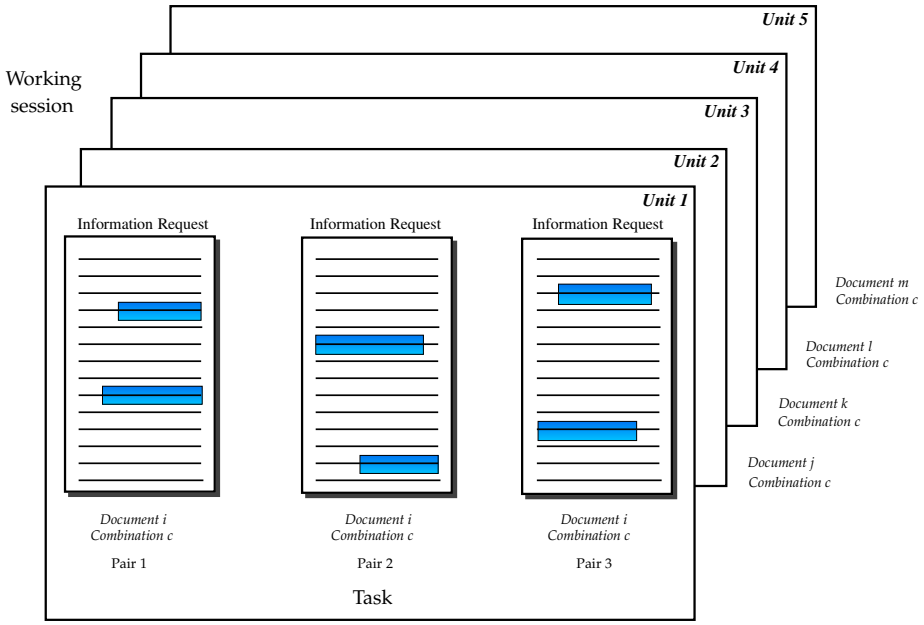

Figure 5.1: Sequence to complete the task in a working session, which is comprised of five documents combinations (units). The information request was placed at the top of the document. The diagram does not explicitly show the gold unit, since this is masked as a normal unit. Note that the three pairs of sentences were displayed sequentially. That is, a participant could first see pair 1, then pair 2, and then pair 3, highlighted in the same document.

5.1.2 Analysis of Indicativeness Assessments

Our study aimed to collect indicativeness assessments of sentences judged as relevant in the Novelty track dataset, since we cannot assume that all sentences judged relevant are helpful for evaluating query-biased summarisation methods. This section details statistics regarding the number of indicativeness judgements collected, and analysis of such assessments.

We gathered 339 indicativeness judgements from 26 trusted workers for the 90 document combinations available. We observed that we collected an unbalanced number of assessments of document combinations. For example, sentence pairs in DC_1, DC_2 and DC_3 collected 4,5 and 6 assessments from different workers, respectively. To have a balanced number of judgements for all document combinations, from these 339 assessments we randomly sampled three judgements for each of the three document combinations. We gathered 268 indicativeness assessment by taking an uniform sample of nine judgements per document, meaning that two out of the 30 documents only had eight assessments in the three combinations. Given the distribution of units in CrowdFlower, we observed that some documents were assessed by the same participant twice. In order to remove duplicate judgements, we took the first assessment done by a worker. For this reason, two documents gathered eight assessments in the three document combinations.

We calculated the proportion of selection (indicativeness) where subjects chose a relevant sentence as also being indicative for a summary. We called this the proportion of selection among workers, defined as:

$$A_w = \frac{\textit{Indicativeness Judgements}}{\textit{Total Judgements}} \tag{5.1}$$

where *Indicativeness Judgements* corresponds to the number of workers that selected the relevant sentence as indicative in a given document, and *Total Judgements* is the number of workers who assessed that sentence. We calculated the overall proportion of selection (A) as shown below:

$$A = \frac{\sum_{d=1}^{D} \sum_{np=1}^{NP} A_w}{D \times NP} \tag{5.2}$$

where D is the total number of documents, and NP is the number of pairs to be assessed per document. That is, $D = 30$ and $NP = 3$.

The box labelled Random in Figure 5.2 shows the distribution of values after applying Equation 5.1, on the randomly sampled judgements, as explained previously. A circle within the box indicates the overall proportion of selection A, as defined in Equation 5.2. Our results

Assessments collected	Pair 1	Pair 2	Pair 3	Total
1	0	0	1	1
2	1	1	1	3
3	1	1	1	3
4	0	1	1	2
5	0	0	0	0
6	1	0	1	2

Table 5.4: Example of assessments collected in a given document combination.

demonstrate that the Novelty track assessments are useful for conducting evaluation of query-biased summaries to a certain extent. That is, 73% of the time that a sentence is judged as relevant in the Novelty track, it is also indicative for a short query-biased summary. These outcomes not only represent an assessor error rate, which we discuss in the following section, but also shows that Novelty track assessments have to be examined prior to conducting evaluation for tasks out of the scope of the track.

We extended our analysis to quantify upper and lower bounds of proportion of selection from our collected judgements. In order to compute these oracle values, we employed the complete set of assessments (339 in total). For instance, Table 5.4 displays six assessments for a given document combination, where 1 denotes that a relevant sentence in a pair was deemed to be indicative, and 0 otherwise. The upper bound approach gave preference to select the first three assessments where the majority of the relevant sentences were chosen as indicative in a given document combination. In contrast, the lower bound did the same for the first three assessments where a minority of relevant sentences were selected. For the example illustrated in Table 5.4, the upper bound approach initially selected assessments 2 and 3. Due to assessments 4 and 6 having the same number of indicativeness judgements, we completed the sample of three by randomly selecting one of them. The lower bound approach included assessments 5 and 1, while the third assessment was randomly chosen in the same fashion as for the upper bound. Having the set of assessments for upper and bound approaches, we computed A_w and A to obtain oracle values, as detailed in Equations 5.1 and 5.2. The overall upper and lower proportion of selection was 78% and 67%, respectively, see second and third boxes in Figure 5.2 for more details. We provide an example of how to use these oracle values in Section 5.1.4.

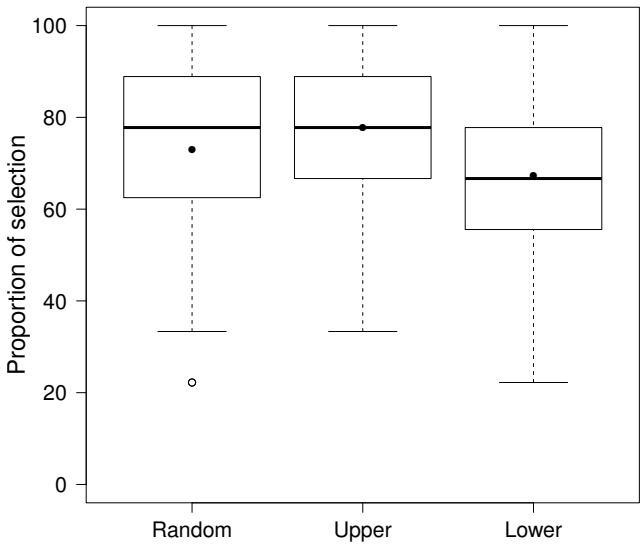

Figure 5.2: Proportion of selection, A_w, of a random sample of judgements, upper and lower bounds of selection. The circle in each box represents the overall proportion of selection, A.

In a more detailed analysis, we quantified the total number of indicativeness assessments where three, two, one or none of the relevant sentences presented per document were chosen as indicative. This analysis was carried out using the random sample of indicativeness assessments (268 in total). Table 5.5 summarises the frequency of selection and the corresponding proportion for three, two, one or none relevant sentences selected as indicative. Given a document combination *DC* with three pairs of sentences, workers frequently selected two and three relevant sentences as indicative. These results are another view that suggests that sentence assessments of the Novelty track do not fully fit to evaluate sentence ranking methods for query-biased summaries, as three relevant sentences were selected as indicative only 39% of the time. In Section 5.1.4, we explain possible reasons that led workers to not select a relevant sentence to assemble a short summary.

	Sentences selected				
	3	2	1	0	Total
Frequency of selection	97	129	35	7	268
Proportion	36.2%	48.1%	13.1%	2.6%	100.0%

Table 5.5: Frequency of selection and proportion of relevant sentences chosen as indicative in sampled assessments.

5.1.3 Assessor Error Rate Adjustment

We mentioned in the previous section that participants selected relevant sentences as indicative around 73% of the time. Thus, we can assume this value is an assessor error rate (α) to adjust P@2 values reported in Chapter 4. This section presents results of a simulation that aims to provide an estimate of the effectiveness of sentence ranking methods using this assessor error rate. A simple approach to adjust P@2 values of previous results is by directly multiplying them by the 73%. In order to define a formal estimation, we simulated a Bernoulli distribution based on α [Turpin et al., 2009].

The simulation relied on the initial sentence assessments provided by Novelty track judges to create *artificial* indicative assessments. Given a random irrelevant sentence and a random number, a relevant sentence was substituted with an irrelevant sentence in the case that this random number exceeded 0.73. Otherwise, the relevant sentence was not replaced. Table 5.6 shows an example of how artificial assessments were generated for a document containing 22 sentences, where 5 sentences were judged as relevant. These artificial assessments were intended to resemble indicativeness judgements rather than topical relevance.

A total of 1,000 artificial assessments were generated for the set of relevant sentences in each document. We observed that some documents in the Novelty track had more relevant than irrelevant sentences. So for the Bernoulli distribution, this could introduce a bias as there were not enough irrelevant sentences to replace relevant sentences. Thus, we selected documents containing more relevant than irrelevant sentences, or at least the same number. A total of 521 out of 1,120 documents satisfied this restriction for the Novelty track 2003, and 843 out of 1,070 documents for the Novelty track 2004. This subset of documents compiled 3,713 and 5,371 relevant sentences for the Novelty track 2003 and 2004, respectively.

Table 5.7 presents results of averaged P@2 scores for sentence relevance assessments given by the Novelty track judge (see third and fourth columns). The fifth and sixth columns

Position of sentences judged as relevant in document	10	13	14	20	21
Position of random irrelevant sentences in document	15	5	19	11	7
Random number	0.51	0.27	0.76	0.16	0.86
Position of sentences in final artificial assessment	10	13	19	20	7

Table 5.6: Example of how artificial assessments were generated for the document NYT19981206.0217 having 5 relevant sentences and 17 irrelevant sentences.

Sentence ranking method	Query	Novelty track		Scaled at α		Estimated	
		2003	2004	2003	2004	2003	2004
QB	*t*	0.46	0.46	0.34	0.34	0.38	0.38
	t+Rocchio-D	0.58	0.54	0.43	0.39	0.46	0.43
	t+LCA-D	0.58	0.54	0.43	0.39	0.46	0.43
VSM	*t*	0.51	0.47	0.37	0.34	0.41	0.38
	t+Rocchio-D	0.59	0.52	0.43	0.38	0.46	0.42
	t+LCA-D	0.58	0.54	0.42	0.40	0.45	0.43

Table 5.7: P@2 values after applying the assessor error rate, α. There are 521 and 843 documents available in the Novelty track 2003 and 2004, respectively.

indicate a scaled P@2 performance given $\alpha = 0.73$. The last two columns detail results of averaged P@2 over artificial assessments obtained by sampling at rate α. As can be seen, the difference between scaled and estimated P@2 values were small; it varied from 0.03 to 0.06. We suggest that using α can provide an approximation of performance of sentence ranking methods for assembling short query-biased summaries. Results from our stochastic simulations reported significant differences (paired Wilcoxon test, $p < 0.001$) of estimated performance between ranking methods that employed a query baseline t against and expanded query, see last two columns of the table. The percentage increase for using the expansion varies from 10% to 21% using the Novelty track 2003. For the Novelty track 2004, the *QB* and *VSM* methods reported a similar increase of 13% for both types of expansion (*Rocchio-D* and *LCA-D*).

5.1.4 Discussion of Sentence Indicativeness

This section describes a failure analysis to identify possible reasons that led participants to not select a relevant sentence as indicative, and outlines possible shortcomings regarding our task design.

Failure Analysis. We carried out an analysis to investigate reasons why participants sometimes did not select sentences that were judged as relevant as also being indicative. Note that failing to choose these sentences (or disagree) does not suggest that workers were doing their tasks incorrectly, since this possibility was controlled in other ways, as explained in Section 5.1.1. There were several factors that could influence their decisions such as the task, the topic, or the sentence itself. In general, we detected two main possible reasons that led subjects to not choose a relevant sentence as indicative. First, participants were given the description of topics as the information request, which could make it more difficult to complete the task. Second, the interface was designed to automatically display the next sentence pair to be assessed. Hence, if workers accidentally selected an irrelevant sentence this could not be modified. Unfortunately, the effect of this event in the interface could not be quantified. By examining the indicativeness judgements, we provide other reasons of disagreement. For example:

- Workers did not tend to select short relevant sentences (less than 10 words) as indicative. Although short sentences had terms from the information request, this type of sentence may not contribute with indicative content about a document. For example, the relevant sentence *"That is also the case with Egypt Air Flight 990"* has three direct matching terms from the request "Information about the Egyptian Air Flight 990 disaster that occurred in October of 1999". For Novelty track assessors, the sentence was topically related, but in participants' opinion the sentence did not point to information that provided a glance of a document content. Section 5.3 provides details on whether the sentence length affected the selection of indicative sentences.

- Another possible reason that caused participants to not select the relevant sentence as indicative was the co-reference resolution problem. This problem consists of identifying the instance that a pronoun refers to in a text. Given the request "Information about John Glenn's flight on the Shuttle Discovery", consider the relevant sentence *"He said he hoped it would lead to a renewal of widespread public support for space exploration"* and the irrelevant sentence *"The other astronauts, usually overlooked outside the Glenn limelight, are Lt. Col. Steven Lindsey of the Air Force, the pilot; Stephen Robinson, a mechanical engineer; Pedro Dugue, an astronaut-engineer, and Dr. Scott Parazynski and Dr. Chiaki Mukai"* presented to participants. The relevant sentence contains a claim about space exploration done by Glenn. However, the reference *"he"* was resolved in a previous sentence not used in our experiment for assessment. Consequently, for

participants the sentence itself did not provide insights about the information request. This example leads to another observation: we suggest that terms such as *astronauts* and *pilot* could influence participants decisions, as this supplied details to participants about the Shuttle Discovery's crew.

- Finally, workers usually chose irrelevant sentences as indicative if they provided a short definition or a general aspect of the topic. For instance, multiple workers found the following sentence as indicative for a summary about "Safety of irradiated food": *"During irradiation, low-level doses of gamma rays or electron beam irradiation are administered to kill bacteria"*. However, this sentence was not labelled as relevant by the Novelty track assessors.

Task Design. Section 5.1.1 explained the creation of document combinations, where relevant and irrelevant sentences were arranged in different orderings. For example, the relevant sentence R_1 in Table 5.2 alternated displaying times in each document combination. We investigated whether this ordering affected the selection of a sentence judged as relevant to be also indicative. Using the random sample of assessments described in Section 5.1.2, we calculated the proportion of selection and the overall proportion of indicativeness. A one-way ANOVA on the pair ordering (first, second, third) indicated that there were no significant differences ($p = 0.117$) when workers selected a relevant sentence as indicative in any order of presentation. These results demonstrated that we did not introduce ordering effects in our experiment or in our analysis.

We did not include a mechanism to gather qualitative feedback from subjects, thus we did not know with certainty what triggered their choices. In addition, we collected assessments for topics that we assumed could be easy or interesting. It was possible that, depending on the topics, the proportion of agreement may vary. A one-way ANOVA test on the topics showed that the proportion of agreement of relevant sentence as indicative was significantly ($p = 0.019$) affected by the topics. We also considered a trade-off when participants assessing sentences, as we did not impose them to read the entire document for completing the task. In these situations, we propose that the lower bound of the proportion of selection (67%) among workers can be used instead of the overall indicativeness (73%).

5.2 Sentence Length

In this section, we study sentence length of relevance assessments in the Novelty track as another property that has not been explored for evaluating sentence ranking approaches. On close examination of summaries produced in Chapter 4, we observed that ranking methods assisted by query expansion tended to assign higher scores to long sentences. As a consequence, these summaries were longer than those generated by ranking methods that did not use query expansion techniques. A possible reason for this selection is that long sentences are likely to contain not only query terms, but also expansion terms leading to a ranking method to potentially boost the selection of these sentences for a summary.

Previous research has shown that the relevance of a document tends to increase with its length, since long documents may include information related to not only one but several requests [Singhal et al., 1996]. Consequently, users are prone to select these documents as relevant. The document length bias was investigated in early TREC conferences to determine the effectiveness of retrieval systems [Singhal et al., 1996]. It was found that if a system took into account the length of a document in the ranking process, it could outperform those that did not. Further research has also confirmed such an effect [Smucker and Allan, 2005; Losada et al., 2008]. However, this length bias has not previously been investigated for sentence ranking tasks.

By inspecting the length of sentences judged as relevant and irrelevant in the Novelty track dataset, we observed that assessors tended to significantly select long sentences as relevant. Thus, it could be the case that the gain detected for using query expansion in the previous chapter was because ranking methods highly scored long sentences given the length bias in the sentence relevance assessments. This section investigates the following questions.

- We introduce a length component in the sentence ranking methods, which is described in the next section. We hypothesise that the performance of *QB* and *VSM* methods increases by introducing the sentence length component to complement the ranking of sentences, given the length bias of the collection assessments. This is detailed in Section 5.2.2.

- We study whether the length bias affects the performance of sentence ranking methods that employ query expansion approaches as those studied in the previous chapter (*Rocchio-D* terms and *LCA-D* terms). We explain these results in Section 5.2.3.

- In Section 5.2.4, we propose an approach to evaluate sentence ranking methods regard-

Feature	Novelty track 2003	Novelty track 2004
Number of documents	1,250	1,808
Number of relevant sentences	15,557	8,343
Number of non-relevant sentences	24,263	28,628
Mean length of relevant sentence	23.69	24.59
Mean length of non-relevant sentence	13.47	15.26

Table 5.8: Composition of the Novelty track 2003 and 2004. Average sentence length is given in words (including stopwords). Statistics for the Novelty track 2004 were given based on documents which include at least one relevant sentence.

less of the sentence length bias in the Novelty track dataset.

5.2.1 Sentence Length Bias

In this section, we detail the length bias detected in the Novelty track relevance assessments, and explain the heuristic of ranking methods based on the length of sentences.

Length Bias in the Novelty Track Dataset. We examined the number of words in sentences judged as relevant and irrelevant. Section 2.9.1 explains that irrelevant documents were introduced in the Novelty track 2004 to make the task more difficult for participants. However, assessors only provided sentence judgements for relevant documents. The identification numbers of irrelevant documents are not known, since this information is not available in the track data. Thus, a large number of sentences were not assessed. To avoid any effects that this could introduce in our analysis, we discarded documents (hence sentences in these documents) that did not contain at least one relevant sentence.

Outcomes show that relevant sentences contain on average around 24 words, while irrelevant sentences are on average 14 words long. A t-test revealed that the sentence length difference was significant ($p < 0.001$). Figure 5.3 shows the length distribution of sentences judged as relevant and irrelevant in both Novelty tracks 2003 and 2004, with details displayed in Table 5.8.

Sentence Length Component in Ranking Methods. Section 4.2.3 explained that the *QB* and *VSM* methods used the position of a sentence in the document to resolve the ties of sentences scoring the same selection value. That is, the decision factor was how close with respect to the beginning of a document a sentence was located. To investigate the sentence length bias, the *QB* and *VSM* methods were extended using a simple heuristic to

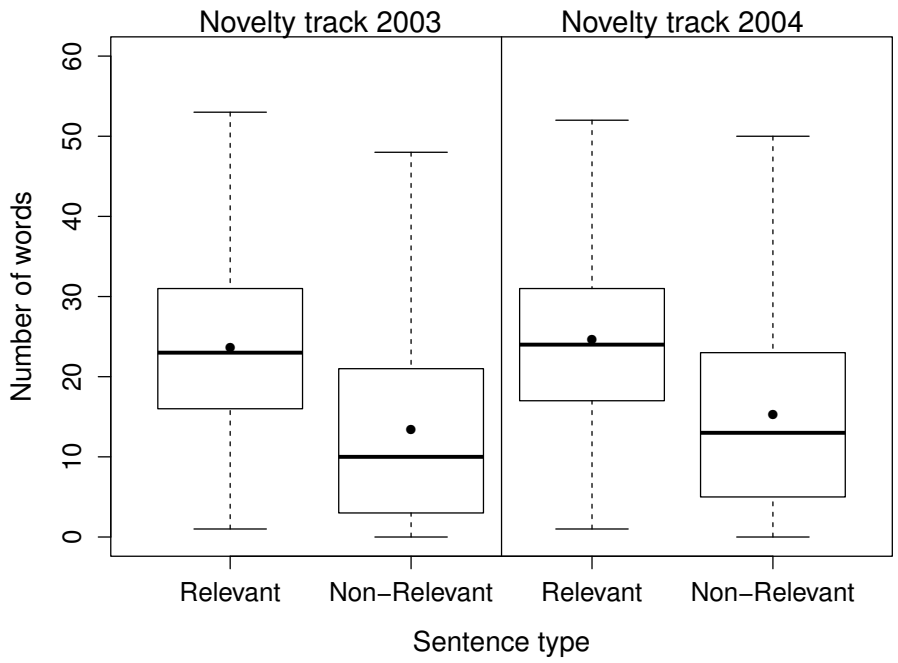

Figure 5.3: Distribution of the number of words in each sentence (including stopwords) in the Novelty track datasets 2003 and 2004, based on statistics reported in Table 5.8.

complement the ranking, based on the length of sentences (measured in words). We identify these methods as *QB-Len* and *VSM-Len*. In this new approach, ties were broken not by the position in the document, but by the length of a sentence, in decreasing order. In other words, long sentences had more priority over short sentences. For example, a ranking method (*QB-Len* or *VSM-Len*) scored four sentences of a document as follows:

Sentence Id	Score	Sentence Length	Final Ranking
1	0.50	12	2
2	0.15	20	4
3	0.35	23	3
4	0.50	17	1

The ranking method used information about sentence length to resolve the ties among sentences 1 and 4, as they scored the same value. The ranking method placed sentence 4 in the first rank position and subsequently sentence 1. The remaining two sentences were ranked according to their score, and the length component was ignored. In the following two sec-

tions we re-examine results obtained in the previous chapter by considering a sentence length component in ranking methods, that is, *QB-Len* and *VSM-Len* methods.

5.2.2 Sentence Length Bias in Sentence Ranking Methods Not Assisted with Query Expansion

Since there is a length bias in the judgements, our first hypothesis tests that using a length component in sentence ranking methods (*QB-Len* and *VSM-Len*) improves the performance over those using a position component (*QB* and *VSM*). We employed documents detailed in Section 4.2 (`sub2003` and `sub2004` datasets), and evaluated the performance of sentence ranking methods based on P@2, as described in Section 4.3.

Table 5.9 outlines P@2 results for the *QB-Len* and *VSM-Len* methods using the title of the Novelty track topic (t) as query, and the title concatenated with the description and narrative (t+d and t+n). These results were compared against P@2 values presented in Table 4.4 of Section 4.4.2. We observed that by simply using the baseline title (t) as query, *QB-Len* or *VSM-Len* methods significantly improved (paired Wilcoxon test $p < 0.001$) over the *QB* or *VSM* approaches that used the position feature. The first and fourth rows in Table 5.10 show that the percentage increase fluctuates between 7% and 26%. For the remaining two baselines (t+d and t+n), *QB-Len* and *VSM-Len* approaches also showed significant differences, with *p-values* detailed in the same table. These results confirm our first hypothesis that ranking methods promoting the inclusion of long sentences outperform those that lack the length component. These findings agree with those reported in the document retrieval context [Losada et al., 2008; Singhal et al., 1996; Smucker and Allan, 2005].

As a point of comparison, we also included a separate *Len* approach (last row of Table 5.9), which exclusively selected the longest sentences, and discarded query terms or other sources of evidence to rank sentences. In Section 4.4.1, we found that the *VSM* method using the title baseline obtained a P@2 score of 0.68, significantly outperforming *QB* (paired Wilcoxon test, $p < 0.001$) for the `sub2003` dataset. In these experiments, the *VSM* method was significantly behind by the *Len* approach (paired Wilcoxon test, $p = 0.022$). This outcome is similar to that found by Metzler and Kanungo [2008], who detected by using machine learning that sentence length was a predominant feature in the Novelty track 2003 to boost the selection of relevant sentences. We noted that in the case of the `sub2004` data, the length effect is more moderated, as the *Len* approach performed similarly to *QB* and *VSM* while employing

Summary method	Query	sub2003	sub2004
QB-Len	*t*	0.77	0.58
	t+d	0.79	0.61
	t+n	0.79	0.63
VSM-Len	*t*	0.75	0.56
	t+d	0.75	0.59
	t+n	0.75	0.62
Len	—	0.72	0.52

Table 5.9: Averaged P@2 over topics for QB-Len and VSM-Len using three baseline queries, while the Len approach is query-independent.

Summarisation method	Query	sub2003		sub2004	
		Δ	p	Δ	p
QB-Len vs *QB*	*t*	26.07%	p<0.001	12.29%	p<0.001
	t+d	19.66%	p<0.001	8.66%	p<0.001
	t+n	7.24%	p<0.001	5.30%	p<0.001
VSM-Len vs *VSM*	*t*	9.78%	p<0.001	6.83%	p<0.001
	t+d	6.32%	p<0.001	2.36%	p=0.007
	t+n	1.31%	p=0.014	1.09%	p=0.121

Table 5.10: Percentage change performance for QB-Len and VSM-Len against QB and VSM, respectively. Significance values are based on a paired Wilcoxon test.

the title as the query.

Results of ranking methods *QB-Len* and *VSM-Len* employing the baseline *t+n* (the title and the narrative fields of a topic) may provide some insights about the expected performance of sentence ranking methods applying formal query expansion approaches. That is, the baseline *t+n* may resemble queries of users being more eloquent to describe their information needs. In fact, we noted that the *QB* method using the *t+n* as query did not improve *QB-Len* employing a single baseline *t*. The P@2 values for the *QB* method were 0.73 and 0.60 for the sub2003 and sub2004 datasets, respectively. As can be seen in the first row of Table 5.9, the P@2 values of *QB-Len* correspond to 0.77 and 0.58 for each corresponding dataset. For the sub2003 dataset, *QB-Len* outperformed significantly the *QB* method by 5% (paired Wilcoxon test $p < 0.001$); however this improvement was not significant for the sub2004 dataset. In the following section, we analyse the effect of the length bias in sentence ranking methods assisted by formal query expansion methods such as *Rocchio-D* and *LCA-D*.

5.2.3 Sentence Length Bias Effect in Sentence Ranking Methods Assisted with Query Expansion

In this section, we investigate whether sentence ranking methods assisted by both query expansion and the length component affects the performance of selecting sentences judged as relevant. Since we demonstrated in Section 4.4.2 that expansion at document level (*Rocchio-D* and *LCA-D*) is effective, we used this approach in our experiments. In the previous section, we briefly discussed that the title of a topic concatenated with the narrative (t+n) barely improved sentence ranking methods over those that only employed the title as baseline query. However, we also noted that such narratives may be too broad to judge the effectiveness of automatic query expansion approaches. For example, a fragment of the narrative of the topic 11 (Hurricane Mitch Central America) states: "*Reports from various countries mentioning Hurricane Mitch's location and strength as it progressed through the Caribbean to Central America were relevant*". While "*from various countries*" is a general description of the information need, automatic expansion approaches detected terms such as `Cayman, island, Honduras` and `Colombia` that may be closely related to the intention of the information request.

We replicated those experiments where the title of a topic was progressively concatenated with batches of five extra terms: either with *Rocchio-D* terms or with *LCA-D* terms. In order to make uniform comparisons with results presented in Section 4.4.2, we also used 45 extra terms as cut-off values. Results shown in the first and second rows of Table 5.11 were computed with this fixed number of *Rocchio-D* and *LCA-D* expansion terms. We also calculated the P@2 performance upper bound that can be expected through the expansion process, with values shown in the third and fourth rows of the same table, and labelled as *Rocchio-D-Best* and *LCA-D-Best*. This calculation employed the optimal number of expansion terms per topic.

Figure 5.4 displays the P@2 performance of sentence ranking methods *QB-Len* and *VSM-Len*, both using the baseline title (t) and expanded queries (t+*Rocchio-D* or t+*LCA-D*). However, as observed in the panels of Figure 5.4, the P@2 performance slightly fluctuated at different cut-off values, and it did not remain stable or flat as in the case of *QB* and *VSM* ranking approaches employing expanded queries. The figure uses circles to denote the performance of *Rocchio-D* expansion, and squares for *LCA-D* expansion. For comparison purposes, panels in the figure include three additional results that have been discussed previously. First, we include the P@2 performance for *QB* and *VSM* using only the baseline title.

		QB-Len		*VSM-Len*	
		sub2003	sub2004	sub2003	sub2004
Expansion	*t+Rocchio-D*	0.77	0.60	0.76	0.58
	t+LCA-D	0.76	0.61	0.75	0.60
Oracle	*t+ Rocchio-D-Best*	0.83	0.67	0.82	0.67
	t+ LCA-D-Best	0.82	0.68	0.81	0.67

Table 5.11: Averaged P@2 over topics of query expansion in sentence ranking methods QB-Len and VSM-Len using the length component.

These ranking approaches employed the position of a sentence to resolve ties, are denoted by a straight line and the legend t,Pos. Second, panels display the P@2 performance for the *QB* and *VSM* methods employing the fixed expansion of 45 extra terms. These results are represented in the figure by the symbol "x" and the legend t,Exp-D,Pos. Third, P@2 values for the *Len* query-independent approach are denoted by a dotted line and the legend Len.

Table 5.12 presents the percentage increase of *QB-Len* and *VSM-Len* methods using 45 terms for expansion against the *t*, *t+d* and *t+n* baselines. Outcomes indicated that adding either *Rocchio-D* or *LCA-D* terms to the baseline title led to negative differences, in particular using the sub2003 dataset. The query expansion process, however, significantly improved around 5% after employing *LCA-D* terms only for the sub2004 data. These results are in bold font in Table 5.12.

Query expansion is a technique that aims to reduce the vocabulary mismatch problem, and to increase the retrieval of useful results. From our findings in this section, one can assume that query expansion did not benefit sentence ranking methods to identify relevant sentences. For example, observe in Table 5.12 that the percentage increases are small and not significant. Thus, the *QB* and *VSM* methods in Chapter 4 were on average not solving the vocabulary mismatch problem, but rather were selecting long sentences. Using the Novelty track relevance assessments and having the length bias in sentence assessments, we cannot clearly determine the effectiveness of query expansion applied to sentence ranking methods. Consequently, we propose two evaluation approaches of sentence ranking methods that control for the length. The first evaluation is explained in the next section, and the second evaluation is a user study detailed in Chapter 6.

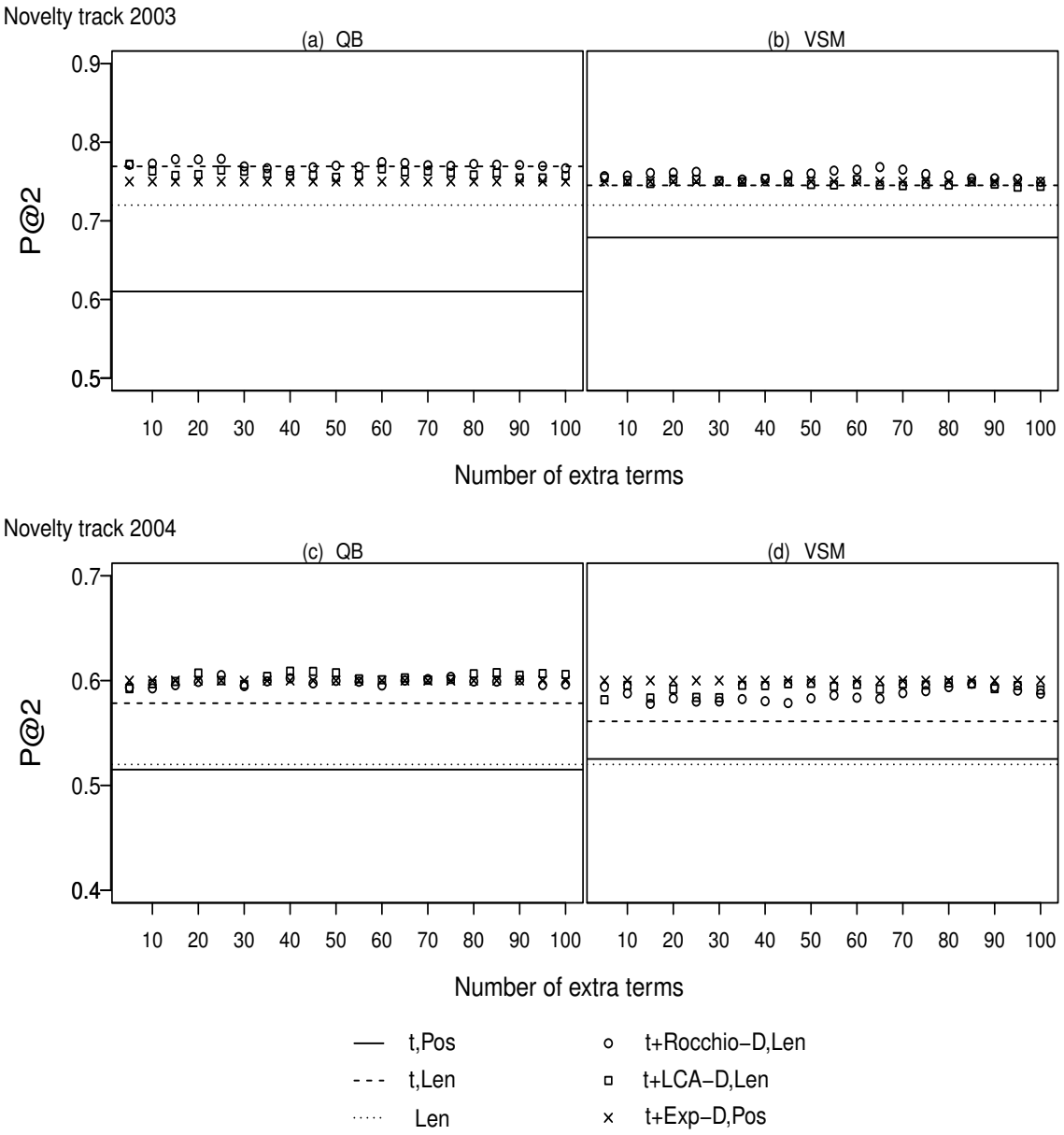

Figure 5.4: Averaged P@2 over topics for QB-Len and VSM-Len using Rocchio-D and LCA-D terms. Panels (a) and (b) use the subset `sub2003` *, and panels (c) and (d) the subset* `sub2004` *.*

Summary type	Expansion vs Baseline	sub2003		sub2004	
		Δ	p	Δ	p
QB-Len	***Rocchio-D***				
	t	-0.15%	p=0.549	3.27%	p=0.172
	t+d	-2.29%	p=0.179	-1.28%	p=0.551
	t+n	-2.41%	p=0.060	-5.32%	p=0.046
	LCA-D				
	t	-1.46%	p=0.506	**5.25%**	**p=0.006**
	t+d	-3.48%	p=0.037	0.62%	p=0.524
	t+n	-3.70%	p=0.301	-3.50%	p=0.492
VSM-Len	***Rocchio-D***				
	t	1.81%	p=0.427	3.11%	p=0.314
	t+d	1.40%	p=0.478	-1.73%	p=0.603
	t+n	0.58%	p=0.527	-7.08%	p=0.013
	LCA-D				
	t	0.63%	p=0.510	**6.39%**	**p=0.005**
	t+d	0.22%	p=0.615	1.40%	p=0.421
	t+n	-0.59%	p=0.183	-4.12%	p=0.390

Table 5.12: Percentage change in P@2 for document-based Rocchio and LCA expansion (at a cut-off of 45 extra terms) using sentence ranking methods QB-Len and VSM-Len. Significance values are from a paired Wilcoxon test.

5.2.4 Isolating the Sentence Length Bias

We propose to isolate the sentence length factor from sentence ranking methods. Our approach resembles that used by Singhal et al. [1996] for document retrieval, which consisted of grouping documents of similar size, measured in bytes. Given a document, we propose to bucket sentences based on the number of words they contain to attempt to isolate the length predisposition. Prior to constructing buckets of sentences, we computed the average word length (μ) of sentences judged as relevant and irrelevant in the Novelty track dataset. In addition to the mean value, we also calculated the word length standard deviation (σ). These values helped to establish a lower and an upper bound, and a middle point to define thresholds for buckets of sentences. We employed relevant and irrelevant sentences outlined in Table 5.8 to calculate the average and standard deviation word length. Computation of μ

Bucket	m	Novelty track 2003	Novelty track 2004	Total
l_s	4	629	588	1,217
l_m	3	722	682	1,404
l_l	2	1,020	983	2,003

Table 5.13: Number of documents in each bucket for the Novelty tracks 2003 and 2004.

and σ values were done separately for the Novelty track 2003 and 2004 datasets. We obtained $\mu = 17$ words and $\sigma = 12$ words for both tracks when rounding to the nearest whole number of words. We defined three sentence lengths for each bucket as follows:

- Short sentences, for bucket l_s, have between 5 ($\mu - \sigma$ is the lower bound) and 13 words;
- Medium sentences, for bucket l_m, contain from 14 to 20 words (μ is the middle point); and
- Long sentences, for bucket l_l, contain between 21 and 29 ($\mu + \sigma$ is the upper bound) words.

Given the sentence lengths in each bucket, the information in a sentence may vary between buckets. For instance, a summary comprised of two long sentences (bucket l_l) could be potentially more informative than a summary containing two short sentences (bucket l_s). In order to minimise such an effect, we modified the number of top-ranked sentences in summaries depending on the sentence length. That is, P@4, P@3 and P@2 were used to report the performance in each of the three buckets l_s, l_m and l_l, respectively. While it has been suggested that short sentences can be simply ignored in the construction of summaries [Kupiec et al., 1995], we argue that these sentences (bucket l_s) can be valuable for space-limited environments such as search result pages. Table 5.13 lists the number of documents in each bucket. This can be seen as splitting the Novelty track dataset into several subcollections, where documents have at least m relevant sentences of length l_i.

This way of grouping sentences enables removing the sentence length bias in the collection. We re-examined the *QB* and *VSM* methods to evaluate the effectiveness of query expansion techniques. In cases where sentences scored the same selection value, ties were resolved by using the position of a sentence within a document. We start by listing the P@2 performance of baseline queries (t, t+d and t+n) in each bucket, with details in Table 5.14.

		Novelty track 2003			Novelty track 2004		
		P@2	P@3	P@4	P@2	P@3	P@4
		l_l	l_m	l_s	l_l	l_m	l_s
Method	Query						
QB	*t*	0.71	0.61	0.31	0.49	0.38	0.19
	t+d	0.72	0.62	0.31	0.50	0.40	0.19
	t+n	0.71	0.62	0.33	0.52	0.40	0.20
	t+Rocchio-D	0.71	0.62	0.33	0.50	0.40	0.20
	t+LCA-D	0.71	0.63	0.34	0.50	0.40	0.21
VSM	*t*	0.71	0.61	0.31	0.50	0.39	0.19
	t+d	0.71	0.62	0.31	0.50	0.40	0.20
	t+n	0.70	0.61	0.33	0.51	0.40	0.20
	t+Rocchio-D	0.71	0.62	0.33	0.49	0.40	0.20
	t+LCA-D	0.70	0.62	0.35	0.49	0.41	0.21
Len	—	0.64	0.57	0.40	0.41	0.34	0.21

Table 5.14: Averaged P@m over topics for buckets of sentences of length l.

		Novelty track 2003			Novelty track 2004		
		P@2	P@3	P@4	P@2	P@3	P@4
		l_l	l_m	l_s	l_l	l_m	l_s
Method	Query						
QB	*t+Rocchio-D*	5	20	10	20	30	35
	t+LCA-D	20	20	75	10	40	50
VSM	*t+Rocchio-D*	10	5	10	5	30	40
	t+LCA-D	40	35	70	15	50	55

Table 5.15: Number of extra terms for optimal expansion for buckets of sentences of length l in each summarisation method.

The results for sentence ranking methods employing expanded queries such as *Rocchio-D* or *LCA-D* terms were based on optimal cut-off values and not on a fixed number of extra terms, as we did in previous experiments that used 45 expansion terms. We observed that the optimal number of expansion terms varied according to the bucket size, sentence ranking method and expansion technique. Table 5.15 lists the number of extra terms required to reach the maximum P@m performance for each bucket.

By analysing buckets of length l_l (long sentences) in both Novelty tracks 2003 and 2004, the expansion did not reveal any improvement over the simple title baseline (t). Likewise the

same behaviour was found for sentences of length l_m. Significant improvements, however, were noticed for sentences of length l_s (paired Wilcoxon test). For documents in the Novelty track 2003, the percentage increase of *QB* and *VSM* methods fluctuated between 7% ($p = 0.007$) and 11% ($p < 0.001$) for *Rocchio-D* and *LCA-D* terms, respectively. In the Novelty track 2004, the percentage change was significant around 10% only when using *LCA-D* terms ($p = 0.006$ for the *QB* method, and $p = 0.002$ for the *VSM* method). Table 5.14 reports values for P@2, P@3 and P@4 performance for each sentence bucket, sentence ranking method and expansion approach.

Figure 5.5 illustrates the gain of query expansion graphically, by grouping sentences of similar length using the *QB* method. The dark gray colour in each bar of the figure denotes the performance of *QB* applying the title of a topic (t). The light gray colour shows the performance of the sentence ranking method after employing query expansion (*LCA-D* terms). The performance values of the *QB* method ignoring the length bias, and using a baseline query are detailed in the first bar of each panel of the figure, and labelled as no buckets. The remaining three bars represent the performance for each bucket of sentences, as defined in this section. Observe in Table 5.14 that the *VSM* method performed similarly to the *QB* ranking approach. Hence, we did not include a figure detailing the gain of expansion for this method. It can be seen that the performance gain from query expansion were over-estimated in the previous no buckets case. When sentence length was accounted for, the gain was minimal for long and medium sentences, and more prevalent for short sentences.

For comparison purposes, we also included the *Len* approach that simply selected the m longest sentences of each bucket. It can be seen in Table 5.14 that for buckets of length l_s, the *Len* approach performed significantly better (paired Wilcoxon test, $p < 0.001$) than using any of the baselines (t, t+d and t+n), and the expansion techniques for the Novelty track 2003 dataset. In contrast, for the Novelty track 2004 dataset, the *Len* method significantly improved only against the baseline title for the bucket l_s ($p = 0.004$ for the *QB* method, and $p = 0.002$ for the *VSM* method). However, we found that sentence ranking methods for buckets l_l and l_m using query expansion significantly improved against the *Len* approach ($p < 0.001$). These results suggest that query expansion effectively assists to resolve the vocabulary gap for a specific sentence length. In Chapter 6, we explore if these findings based on a Cranfield-based evaluation can be confirmed in a user study.

In general, we found that ranking methods barely improved from using baseline queries to expanded queries. By following our proposed methodology, of creating subcollections by bucketing sentences of similar length, we discovered that the expansion was useful to a

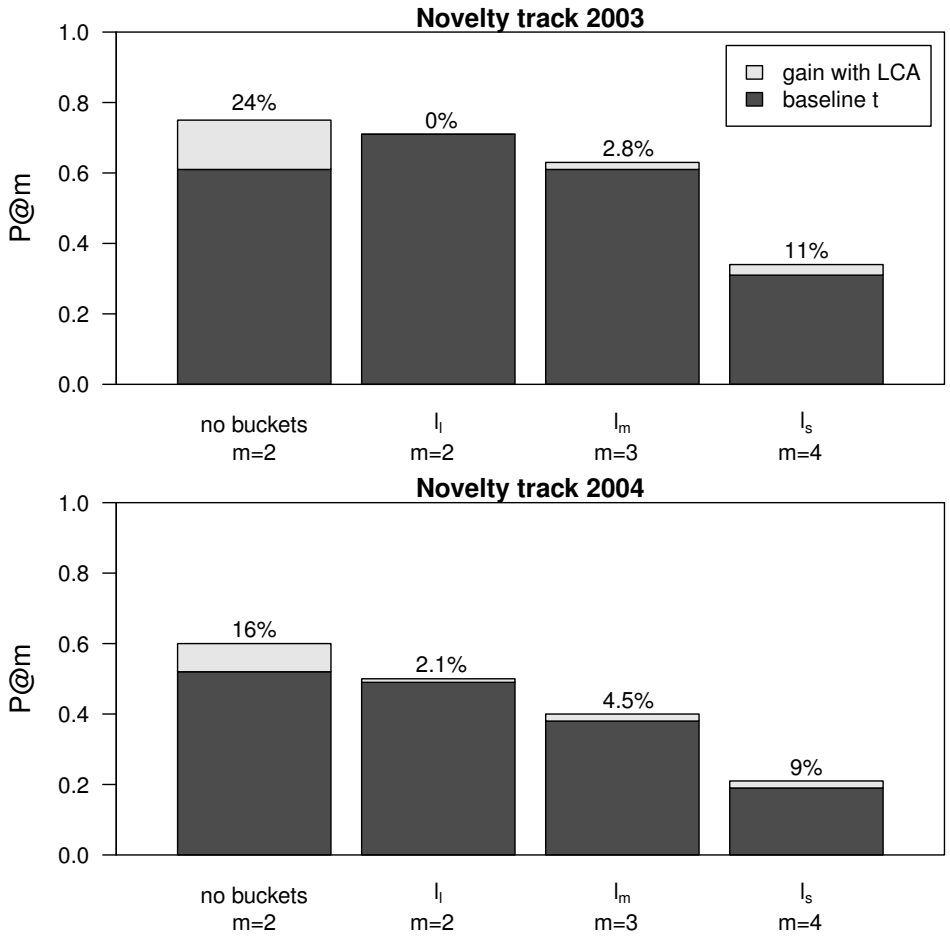

Figure 5.5: Gain of expansion terms over the QB sentence ranking method for different buckets. The percentage at the top of each bar denotes the percentage increase of query expansion approaches (or gain).

very limited extent. That is, the selection of relevant sentences was improved only for short sentences (bucket l_s) having between 5 and 13 words. We observed that this bucket required more expansion terms compared to the other two buckets. A possible explanation to this is that sentences in bucket l_s are shorter compared to the other two buckets. Therefore, the probability of containing expansion terms is also reduced.

In this section, we have found that sentence length is a property of the Novelty track relevance assessments that can mislead conclusions about the effectiveness of sentence ranking methods for constructing query-biased summaries. These results are based on a Cranfield style evaluation. By controlling the length, our outcomes suggest that query expansion significantly improved for short sentences. However, the performance of the *Len* method,

as shown in Table 5.14, poses a question to investigate this approach in detail. In the next chapter, we present results of a user study that compares sentence ranking methods assisted by query expansion against the *Len* approach.

5.3 Discussion

Marcus et al. [1978] discovered that the indicativity (relevance) of bibliographic catalog fields was correlated with their length. That is, long fields such as a set of phrases created from the document vocabulary and abstracts tended to be more indicative than titles of documents and phrases matching search terms. We inspected the description of catalog fields studied by Marcus et al. [1978], and found that few fields may provide information using only one sentence. Thus, we re-used the assessments collected in our crowdsourcing experiment (explained in Section 5.1) to investigate any relation between sentence length and sentence indicativeness. In this section, we analyse relevant and irrelevant sentences exclusively employed in our crowdsourcing experiment, and not to the entire set of relevant and irrelevant sentences of the Novelty track dataset. In the study, we paired sentences according to their relevance status given by the Novelty track assessor rather than any other particular feature. This analysis is two-fold.

- We investigate whether sentence length affects the selection of a relevant sentence as indicative. This is detailed in Section 5.3.1.
- For each bucket of sentences of a particular length defined in Section 5.2.4, we identify the proportion of selection (indicativeness) based on judgements collected in the user study detailed in Section 5.1, and how this affects the perceived performance of sentence ranking methods. We present this analysis in Section 5.3.2.

5.3.1 Sentence Length of Indicativeness Judgements

Recall that to gather indicative judgements, participants were presented with a relevant and irrelevant sentence within a document, and their task was to select the one they considered a good candidate for including in a short summary. We carried out an ANOVA test to investigate whether the length bias affected workers' perception to determine indicative sentences. We found that the sentence length was a factor that significantly ($p < 0.001$) influenced those relevant sentences also deemed to be indicative. This outcome can be related to the findings of Kaisser et al. [2008], where subjects preferred long answers in the form of paragraphs

or sentences to fulfill their information requests that required to explain cause, effect or to describe a process. Jing et al. [1998] also detected that assessors chose long sentences for generic summaries. Previous research has found that the length of a document [Losada et al., 2008; Singhal et al., 1996; Smucker and Allan, 2005] or the length of catalog fields [Marcus et al., 1978] is a factor to determine the relevance of documents. We showed earlier in this section that long sentences tended to be topically relevant. With our analysis of collected indicativeness assessments, this is also the case for sentences. We can conclude that sentence length is also a factor to determine if a sentence is indicative.

In order to quantify the proportion of selection of indicative sentences regardless of the length bias, we split relevant sentences into three groups, based on the length in words of relevant (R) and irrelevant (I) sentences.

- The first group, $R = I$, denoting that the absolute difference in length between relevant and irrelevant sentences is $\leq$5 words;
- the second group, $R > I$, showing that the length of relevant sentences is 5 words or more longer than irrelevant sentences; and
- the third group, $R < I$, showing that the length of relevant sentences is 5 words or more shorter than irrelevant sentences.

We assumed that a difference above 5 words could be visually perceived by a subject. Figure 5.6 displays the proportion of selection of relevant sentences that are indicative in each group. To visualize the sentence length difference, the plot was segmented into three areas that corresponded to the groups defined above. Data points located in the middle panel (between the dotted lines) present the proportion of selection of the group $R = I$, while data points located in the left and right panels show the data for the groups $R < I$ and $R > I$, respectively. The boxplot on the right outlines the proportion of selection of relevant sentences that were also indicative in each group. Note that information displayed in the figure was computed using Equations 5.1 and 5.2 of Section 5.1.2.

We focus on results obtained for the group $R = I$, since we aim to study indicative assessments without the sentence length predisposition. As can be seen in the second box of the figure, assessors selected relevant sentences regardless of their length approximately 74% of the time. An ANOVA test on the proportion of selection as indicative demonstrated that relevant sentences in the set $R = I$ were frequently selected as indicative when the sentence length feature was isolated ($p < 0.001$).

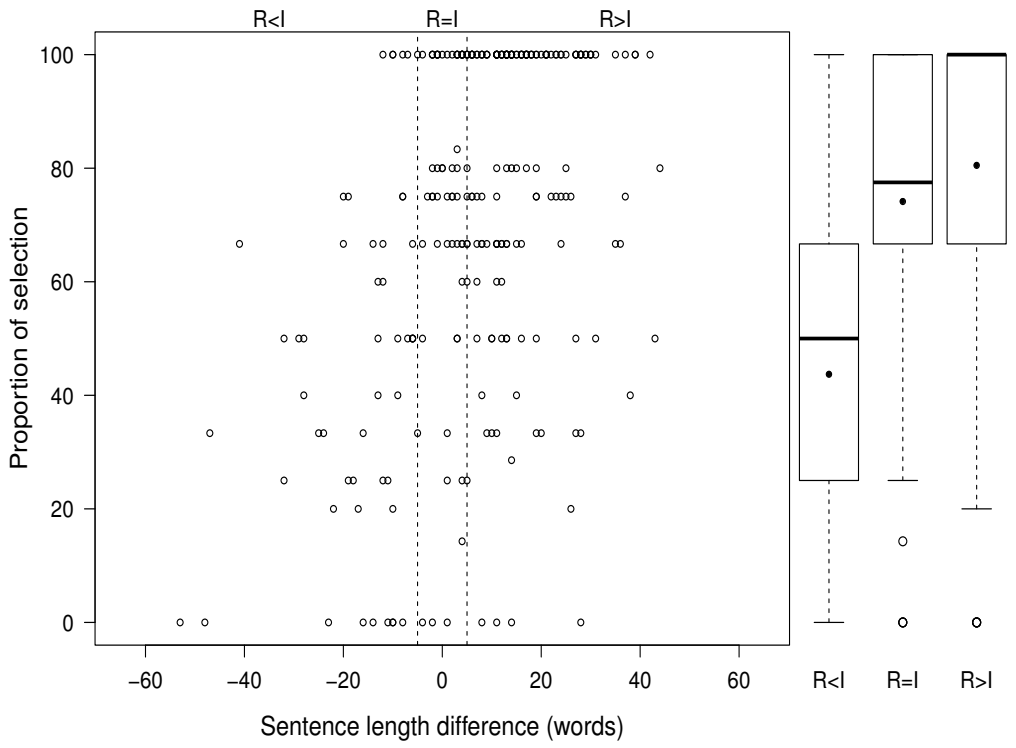

Figure 5.6: Length difference of relevant sentences and irrelevant sentences in terms of number of words. The X-axis shows the positive or negative difference with respect to the relevant sentence, while the Y-axis represents the proportion of selection of the relevant sentence as indicative.

Regarding the other two groups of sentences, we noted that for the $R > I$ group, relevant sentences were frequently selected as indicative. In contrast, for the $R < I$ group, participants tended to disagree more often with the Novelty track assessments when judging short relevant sentences. The following example shows a relevant sentence, which was not selected as indicative for the topic "cloning Dolly the sheep". However, the long irrelevant sentence was often considered indicative by participants.

Relevant sentence:	*But with Dolly, cloning seemed almost impossibly arduous*
Irrelevant sentence:	*"All the other cloning announcements were from biotech companies or agricultural scientists," said Dr. Lee Silver, a molecular geneticist at Princeton University*

Our results showed that participants were influenced by the sentence length to indicate

whether a sentence was indicative or not. Nevertheless, recall that a summary is assembled with the top m sentences. Thus, a user makes a judgement based on a compilation of sentences, rather than single sentences as we have explored in this section. We investigate this in a user study (an extrinsic summarisation evaluation approach) in the next chapter.

5.3.2 Sentence Indicativeness Relative to Short, Medium and Long Sentences

Our analysis in Section 5.2.4 controlled sentence length to investigate the effectiveness of sentence ranking methods given the relevance assessments of the Novelty track dataset. In this section, we aim to join both properties, indicativeness and length, to provide a closer estimate of sentence ranking approaches assisted by query expansion. In the previous section, we observed that the proportion of selection among participants that a sentence judged as relevant was also indicative regardless of the length was around 74% of the time. However, this finding included sentences of different lengths.

We re-used the assessments collected in the study detailed in Section 5.1, and classified sentences in three buckets: short (l_s); medium (l_m); and long (l_l) length. Recall that we defined the length for each sentence bucket in Section 5.2.4. Table 5.16 presents the average proportion of selection for each sentence group. The proportion of selection for each sentence was calculated as defined in Equation 5.1 in Section 5.1.2. We then averaged these values according to the number of sentences available in each group (see third column in Table 5.16). Note that this table includes an extra group, where the sentence length exceeded 30 words but had less than 47 words. We explained at the beginning of this section that we randomly selected sentences for our study, so this explains the extra group, with lengths that did not occur in the analysis of Section 5.2.4.

From results in Table 5.16, one could say that the average indicativeness seemed to rise as the sentence length was increased [Marcus et al., 1978]. This applied for sentence buckets l_s, l_m and l_l; however, this was not the case for the Longer sentences group. In this group the proportion of selection was slightly lower than for the sentence bucket l_l, where the maximum length was 29 words. Figure 5.7 illustrates the proportion of selection in each sentence group.

Sentence bucket	Sentence length (words)	Number of sentences	Indicativeness
l_s (short)	5-13	12	0.58
l_m (medium)	14-20	26	0.68
l_l (long)	21-29	24	0.80
Longer	30-46	27	0.74

Table 5.16: Indicativeness (average proportion of selection) for each sentence group.

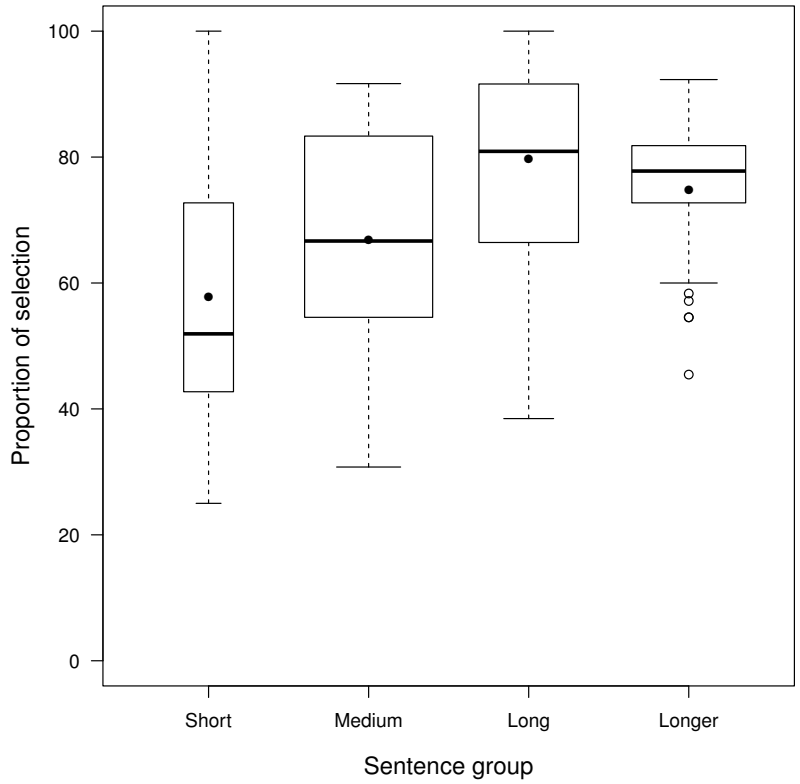

Figure 5.7: Indicativeness (proportion of selection) for each sentence group. The mean is identified with a dot in each box. The width of each box varies according to the sentence population in each bucket, as detailed in Table 5.16.

The proportion of selection in each bucket can be then used to adjust the performance of sentence ranking methods, as described in Section 5.1.3. The first row of Table 5.17 shows P@2, P@3 and P@4 for long, medium and short sentences, respectively. These results were given in terms of topical relevance (see Table 5.14 of Section 5.2.4). Since there were no significant differences between sentence ranking methods and query expansion techniques,

Assessments	Novelty track 2003			Novelty track 2004		
	P@2	P@3	P@4	P@2	P@3	P@4
	l_l	l_m	l_s	l_l	l_m	l_s
Relevance only	0.71	0.63	0.34	0.50	0.40	0.21
Indicativeness-adjusted	0.57	0.43	0.20	0.40	0.27	0.12

Table 5.17: Performance for buckets of sentences of length l.

we focused on the results of the method *QB* using *LCA-D* terms, presented in the first row of Table 5.17. In the second row of the same table, we show the estimated values of P@2, P@3 and P@4 after adjusting for the sentence indicativeness for each group. For example, P@2 for the bucket of long sentences in the Novelty track 2004 was 0.50. Thus, the estimated P@2 is 0.40 after applying the adjustment for long sentences (0.80) for indicativeness. These results agreed to those presented in Section 5.1 that the effectiveness decreased on an indicativeness basis. In addition, the estimated effectiveness of sentence ranking methods is the same using an unexpanded query or an expanded query as shown in Table 5.14. We extend our evaluation of sentence ranking methods employing query expansion in a user study detailed in the next chapter.

5.4 Summary

TREC launched the Novelty track from 2002 to 2004 to study filtering methods to identify relevant information and novelty detection. The track provided relevance assessments at the sentence level, which led the research community to adapt these assessments to evaluate applications such as snippet generation [Metzler and Kanungo, 2008] or passage retrieval [Losada, 2010]. However, no previous research has been conducted regarding the properties of sentence relevance assessments in the Novelty track. In particular, we studied sentence indicativeness and sentence length. Experiments conducted in this chapter demonstrated that both properties affected the perceived effectiveness of sentence ranking methods assisted by query expansion.

We discover that around 73% of the time participants agree that a relevant sentence is also indicative and, that therefore it can be considered as a good candidate to be included in a short query-biased summary. This result suggests that effectiveness of sentence ranking methods cannot solely depend on assessments of topical relevance as those available in the Novelty track. Through a stochastic simulation, we found that query expansion techniques

significantly assisted sentence ranking methods between 13% and 21%.

We observed that sentences judged as relevant by Novelty track assessors tended to be longer, which led us to examine the effects of sentence length in ranking methods. By including a length component in sentence ranking methods, we could not quantify the effectiveness of query expansion to solve the vocabulary gap problem. That is, using assessments of the Novelty track, a method that favours the selection of long sentence will represent a gain. We proposed an evaluation approach that isolated the sentence length to explore improvements of expansion approaches, which were masked through the biased assessments in the collection. This length-controlled approach revealed that ranking methods assisted by query expansion improved significantly for short sentences (5-13 words). We also provided an estimate of the effectiveness of sentence ranking methods using query expansion techniques when sentence indicativeness and length were considered during the evaluation.

A Cranfield-based evaluation suggests that sentence ranking methods barely improve while using query expansion approaches. In addition, evidence from our collected indicativeness judgements points out that users tended to select long sentences. These facts raise two questions: whether query expansion is effective for assembling query-biased summaries and solving the vocabulary mismatch problem; and whether a *Len* approach, which relies only on selecting long sentences, can effectively substitute for more complicated expansion approaches. These questions are addressed in the next chapter.

Chapter 6

Extrinsic Summary Evaluation

Methodologies for the evaluation of summaries can be classified in two types: intrinsic and extrinsic, as explained in Sections 2.8.1 and 2.8.2. Intrinsic methodologies aim to measure vocabulary overlap between a model and an automatic summary. Extrinsic methodologies, on the other hand, require subjects to assess summaries while performing specific tasks. We have studied several intrinsic methodologies to evaluate query-biased summaries. Chapter 3 introduced a position-based evaluation, which relied on eye tracking data. In Chapters 4 and 5, we showed results of a large scale intrinsic evaluation using assessments of the Novelty track to gauge the effectiveness of query expansion techniques applied to sentence ranking methods. We found that by ignoring properties of the Novelty track assessments such as sentence indicativeness and sentence length, query expansion significantly improved sentence ranking methods. However, we discovered that performance in terms of P@2 for sentence ranking methods was significantly affected by these two properties.

In this chapter we present the results of an extrinsic evaluation to investigate whether users prefer summaries assisted by query expansion. This is related with our third research question. Previous work has shown that summaries biased towards information requests guide users to detect potential relevant documents [Tombros and Sanderson, 1998; White et al., 2003]. Thus, query expansion may be a supportive technique for the generation of query-biased summaries. We know from our previous experiments that sentence ranking methods employing query expansion tended to select long sentences for inclusion in a summary: this leads to increase the length of a summary.

Long summaries can contain more information, appear more coherent and may increase readability [Jing et al., 1998]. These characteristics may help a user to potentially make a

more accurate decision about reading a document in full. On the other hand, users may be overwhelmed by the amount of information displayed in a long summary, and thus prefer a succinct compendium of a document. We employ our novel approach (discussed in Section 5.2.4 of the previous chapter) for constructing query-biased summaries through query expansion with a uniform number of sentences, where sentences have a similar length. In this way, we aim to avoid the introduction of a summary length bias to assessors, thus leading to a more reliable extrinsic evaluation.

6.1 Controlling Length in Summary Evaluation

We identify different approaches to measure the length of summaries in both intrinsic and extrinsic methodologies for summary evaluation, based on previous research. These include, in order of complexity: bytes, characters, lines, words, clauses, sentences and paragraphs. In 2004, the DUC conference restricted the size of single-document summaries to a maximum of 75 bytes, and of multi-document summaries to 665 bytes. When participant systems generated longer summaries, they were truncated to the byte-sizes described above. In search result pages, the space to display summaries can be restricted to a certain amount of characters. For example, the INEX Snippet Retrieval track evaluated summaries of at most 300 characters [Trappett et al., 2011]. Cutrell and Guan [2007] modified the length of snippets shown by a commercial search engine. Based on eye tracking evidence, their results showed that users found long snippets (6-7 lines) to be more helpful for informational requests. That is, requests that require specific information that can be gathered through the inspection of several documents. Generally, DUC/TAC conferences restrict summary length to words. The number of words varies according to the summary type from 10-400 words as shown in Table B in Appendix B. In the literature we found that measuring summary size in terms of clauses is scarce [Jing et al., 1998], possibly as clauses may be too short to express a complete idea.

We note that using bytes, characters, lines or words to quantify the amount of information in a summary may introduce shortcomings, such as lack of coherence and readability issues [Jing et al., 1998; Kanungo and Orr, 2009], since sentences can be truncated. Another way to restrict summary length is by using paragraphs [Salton et al., 1997]. Nevertheless, paragraph-length may not be appropriate for constructing short query-biased summaries. Therefore, sentences can be more suitable to limit the length of a summary, and to exclude possible readability or coherence problems.

In order to avoid the summary length bias, and to conduct a more reliable evaluation, others have employed summaries with the same number of sentences [Jing et al., 1998; Varadarajan and Hristidis, 2006]. We observe that sentence length is an aspect that has been ignored in previous evaluation approaches, however. Using the Novelty track assessments, we demonstrated in Chapter 5 that sentence length plays an important role to gauge sentence ranking approaches. We argue that our previous findings can serve to design robust extrinsic evaluation approaches. We conduct a series of crowdsourcing experiments to investigate the following particular aims related to our third reseach question.

- We investigate whether subjects prefer summaries assisted by query expansion when the length of the summary can be regarded as similar to summaries that do not employ query expansion. That is, summaries using or not using the expansion are comprised of a uniform number of sentences, where each sentence has a similar number of words. This removes a potential bias in assessments towards the amount of information displayed in summaries. We detail these results in Section 6.3.

- In the previous chapter, the *Len* method significantly improved the selection of short sentences. This method simply chooses the longest sentences for a given sentence bucket (long, medium or short sentences). To confirm whether the expansion is effective regardless the length, we analyse if participants prefer a summary constructed using query expansion or employing the *Len* approach. This is discussed in Section 6.4.

These experiments enable us to explore if query expansion techniques help in reducing the vocabulary gap, since our previous Cranfield-based evaluation (based on relevance data) indicated that the expansion improved the selection only for short sentences. We describe our experimental setting in the next section.

6.2 Experimental Setting

We have explained in Section 2.8 that extrinsic summary evaluation can be divided in pseudo-purpose and full-purpose. In this chapter we followed a pseudo-purpose methodology, where participants were presented with two summaries and their task consisted of selecting the summary that would trigger their decisions to inspect the underlying document. In the first experiment, one of the summaries employed expansion techniques to rank sentences, while the other summary used a short query to rank sentences. That is, participants provided judgements regarding whether they prefer a summary aided by query expansion. In the

second experiment, one of the summaries was assisted by query expansion and the other summary was created by selecting the longest sentences. This section details the summarisation approach, sample size, and task provided to participants.

Full-purpose methodologies evaluate how summaries support users to complete a specific task. In the case of query-biased summaries, a simulated task may consist of users identifying relevant documents given a ranked list of results. Since documents of the Novelty track have been previously judged as relevant by TREC assessors, we assumed that full-purpose methodologies were out of the scope in this regard. In future work, documents of the Novelty track 2004 can be used to replicate a full-purpose approach as this collection involves relevant and irrelevant documents.

6.2.1 Summarisation Approach

We found in Section 5.2 that long sentences were more likely to be selected by sentence ranking methods that used query expansion. In addition, Novelty track assessors tended to select long sentences more frequently as relevant. Given these shortcomings, we proposed to bucket sentences according to their length and to evaluate effectiveness at different P@m values. That is, we measured P@2, P@3 and P@4 for summaries comprising long, medium and short sentences, respectively. Note that m denotes the number of top-sentences used to assemble a summary, and according to this value is the number of words allowed in sentences. Long sentences contain between 21 and 29 words, medium sentences from 14 to 20 words, and short sentences have between 5 and 13 words. These ranges were defined in Section 5.2.4 of the previous chapter. By measuring P@2, P@3 and P@4, we found that the *QB* method in combination with *LCA-D* terms for expansion was slightly better than *VSM* (see Table 5.14). Nevertheless, differences between both methods were not statistically significant (paired Wilcoxon test, $p > 0.050$), excepting for buckets of short sentences for the Novelty track 2003. Therefore, in this chapter, we adopted the *QB* method to assemble summaries and *LCA-D* as the query expansion approach. Recall that the *QB* method, explained in Section 4.2.3 ranks sentences according to the occurrence of query terms in a sentence. The number of *LCA-D* extra terms employed for each number of sentences (m) in a summary were listed in Table 5.15.

In order to create baseline summaries for our first experiment, the *QB* method used the title of Novelty track topics as query to rank sentences. In contrast, the other summaries employed the same sentence ranking approach and *LCA-D* terms to expand the title

(*t+LCA-D*). We identify summaries produced with a simple query baseline as *QB-B*, while *QB-E* are summaries constructed with an expanded query. Specifically, the *QB-B* or *QB-E* methods returned the top m sentences, which were then concatenated in the order they appeared in the text and presented as summaries. For our second experiment, the *Len* method chose long sentences (given a sentence bucket) and ignored any other source of evidence.

6.2.2 Sample Size

Prior to the study, we conducted a power analysis to determine the number of document summaries required to test our hypotheses of this chapter — whether subjects prefer summaries assisted by query expansion techniques. To estimate the document sample size, we used results from Section 5.2.4 for summaries comprising long, medium and short sentences. We observed two typical scenarios that occurred when computing P@m in a sentence ranking context:

- First, given a document, the *QB-B* and *QB-E* ranking methods may select the same sentences and therefore equal P@m scores. Thus, such documents should not be considered for computing the sample size. For a given m top ranked sentences, Table 6.1 lists the total of documents discarded in our analysis in parentheses (), as their summaries contained exactly the same sentences. Consequently, we employed documents where the *QB-B* and *QB-E* summaries had $0 \leq m \leq m-1$ common sentences. For example, 595 documents were ignored for $m = 3$, while 809 documents were used for subsequent analysis, as they had either 0, 1 or 2 summary sentences in common. Table 6.1 details the total of documents according to the number sentences in common for the two summarisation approaches when $m = 2$, 3, or 4. The first row in Table 6.2 presents the number of documents without m sentences in common.

- Second, *QB-B* and *QB-E* methods can generate summaries for a given document with different sentences, and score same P@m values (tied scores) or different P@m values. Employing outcomes from documents without m sentences in common, the power analysis resulted in a large sample size that exceeded the number of documents available. To minimise such an effect, from these documents we removed documents where *QB-B* and *QB-E* summaries obtained a tied P@m score. The total number of documents without m common sentences and without tie scores are shown in the second row of Table 6.2. Using these documents, we then carried out a Wilcoxon signed rank test

Common sentences in summary	Documents m =2	m =3	m =4
0	135	16	4
1	807	191	35
2	(1,061)	602	205
3	- -	(595)	520
4	- -	- -	(453)
Total documents	2,003	1,404	1,217

Table 6.1: Number of documents with common sentences given a specific m cut-off after using QB-B and QB-E. The symbol - - denotes that there are no sentences available for that m value.

	m =2	m =3	m =4
Documents without m common sentences	942	809	764
Documents without m common sentences and removed P@m tied scores	321	281	278

Table 6.2: Number of documents with common sentences and with tied P@m scores.

(matched pairs) to calculate the sample size employing the statistical tool G*Power 3.[1] G*Power 3 includes a wide range of statistical power analysis for different tests types.

The sample size gave 224 and 22 documents for summaries containing three and four sentences, respectively, with a statistical power of 80%. According to Cohen [1988], the effect size observed in our analysis is small ($d = 0.17$) for m=3 and moderate ($d = 0.57$) for m=4. For m=2, the power analysis indicated that we required 931 out of only 321 documents available. Moreover, the effect size was very small ($d = 0.08$). Hence, we discarded summaries containing two sentences ($m = 2$) for the experiment. That is, summaries comprised of two long sentences (21-29 words) were not included for evaluation.

Before selecting the samples, we counted the number of common sentences in *QB-B* and *QB-E* summaries from the set of documents where tied P@m scores were removed. Details are given in second and third column of Table 6.3 for $m = 3$ and $m = 4$, respectively. We therefore gathered a random stratified sample size for documents according to the number of common sentences. For example, we counted that for 94 document summaries comprised

[1] `http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/`

Common sentences in summary	Documents m=3	Documents m=4	Sampled documents m=3	Sampled documents m=4
0	10	3	8	- -
1	94	19	74	2
2	177	95	142	8
3	- -	161	- -	13
4	- -	- -	- -	- -
Total documents	281	278	224	23

Table 6.3: Number of common sentences in the sample and stratified sample for $m = 3$ and $m = 4$.

of three sentences (m=3), the *QB-B* and *QB-E* summaries share one sentence in common. That is 33% of the documents available for m=3. This represents 74 documents from our total sample size (224 documents). The last two columns in the same table list the number of documents we employed in our experiment. Note that for summaries comprised of four sentences, the *QB-B* and *QB-E* methods generated different summaries for three documents (none sentences in common between summaries). Since the stratified sample was very small, we did not include documents for this rubric.

In order to test our second hypothesis — whether subjects prefer summaries assisted by query expansion techniques against a summary assembled with long sentences — we re-used the sample size defined above. However, we observed that 13 documents had the same summaries. That is, the *Len* and *QB-E* methods selected the same sentences in a document to create a summary. Consequently, these documents were substituted by randomly selecting other 13 documents so the summaries were not the same.

6.2.3 Task

Our user study aimed to investigate whether users find summaries that have been created using query expansion techniques were more supportive to indicate relevant documents. We collected participants judgements through the crowdsourcing platform CrowdFlower, explained in Section 2.10.3. Participants were presented with an information request and a pair of summaries generated from the same document. We asked subjects to choose the summary that they considered more helpful to decide whether to read the underlying document in full. Recall from section 6.2.1 that a baseline summary simply employed the title of a topic

Instructions: Two systems have created summaries of the same document, these summaries are intended to be representative of the content. You will receive an information request (located at the top of each unit) and your task is to decide: **which summary would be more helpful in deciding whether to go on and read the underlying document in full according to the request?** You have to select one summary: "Summary A" OR "Summary B" To complete your task, you also have to provide brief reasons for your selection. Be careful as requests may not be the same among units.

Table 6.4: Instructions provided to workers in the crowdsourcing task.

to rank sentences (*QB-B*), while an expanded summary (*QB-E*) concatenated the title with *LCA-D* terms. We rotated the order in which the baseline and expanded summaries were displayed, so participants were not able to detect them beforehand. That is, summary pairs took two arrangements (*QB-B*, *QB-E*) or (*QB-E*, *QB-B*) in the first experiment. For the second experiment, the pairs were as follows: (*Len*, *QB-E*) or (*QB-E*, *Len*).

Given that the task was relatively straightforward to complete, workers were asked to provide also a short explanation justifying their selection. A textarea below the summary pairs enabled participants to record their explanation. We present the full participants' instructions in Table 6.4, and a screen-shot of the interface is shown in Figure 6.1.

A unit of work in the CrowdFlower platform consisted of assessing a pair of summaries and explaining the reasons that triggered their selection. In Section 2.10.3, we mentioned that CrowdFlower suggests to employ as gold units around 5% and 10% additional to the total number of work units. We created 22 and 3 gold units for summaries comprising three and four sentences, respectively, for the total sample size detailed in Table 6.3. We requested a total of six judgements for each document's summary pair, with three in each ordering configuration: (*QB-B*, *QB-E*) and (*QB-E*, *QB-B*). Participants could assess up to 25 pairs of summaries, distributed in working sessions of five units. This restriction allowed us to include a large number of participants. We collected judgements separately for summaries with three and four sentences. In other words, once subjects accepted to participate in our experiment, they assessed summaries with the same number of sentences, with either three or four sentences in all their working sessions. We followed the same considerations as detailed

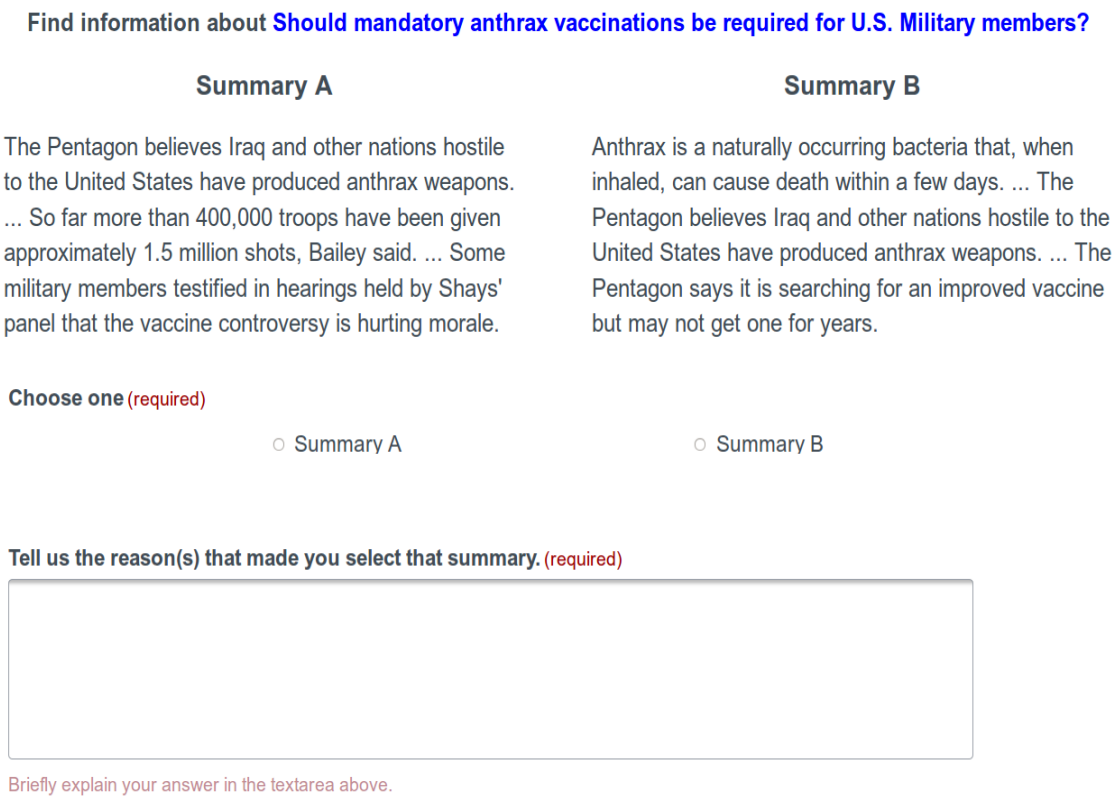

Figure 6.1: Interface snapshot of the crowdsourcing experiment. The information request is located above of each pair of summaries.

above for gathering judgments for the pairs (*Len*, *QB-E*) and (*QB-E*, *Len*).

6.3 Results of *QB-B* vs *QB-E*

We designed a task to collect preferences of users towards a summary that did not employ query expansion (*QB-B*) and a summary that did employ expansion (*QB-E*). In this section we provide details of collected data and analysis of results.

6.3.1 Collected Data of *QB-B* vs *QB-E*

As mentioned before, the task was relatively easy to complete, so we asked workers to provide a short explanation justifying their selection. This mechanism allowed us to ignore judgements of workers who answered gold units correctly, but their explanation was out of the context of the task. This fact suggested that participants passed gold units by chance

without paying attention to the task. We manually inspected the feedback of each worker and, if it contained an explanation out of the context of the task or information request, it was discarded for analysis. For example, workers that wrote non-sense text had their answers removed or such as typed single characters in the textarea for feedback. Given the setting of our task, *trusted* workers were those who successfully answered gold units and provided feedback according to the task. We collected 1,344 judgments from 73 trusted workers for summaries having three sentences, while 138 judgments came from 11 trusted workers for summaries with four sentences. This compiled six assessments for 224 document summaries of length m=3, and six for 23 document summaries of length m=4, the sample sizes detailed in Table 6.3.

6.3.2 Analysis of *QB-B* vs *QB-E*

Table 6.5 displays the number of workers who preferred either the baseline *QB-B* or the expanded *QB-E* summaries. The first part of the table outlines participants' preferences of summaries comprising three sentences of medium length (14-20 words); similarly, the second part does the same for summaries of four sentences of short length (5-13 words). Table 6.5 also lists the number of preferences for common sentences between summaries. For instance, consider summaries comprised of three sentences (m=3) with one sentence in common: 162 workers selected *QB-B* summaries, while 282 preferred *QB-E* summaries. The symbol - - indicates there are no documents available for that amount of common sentences according to the sample size detailed in Table 6.3. We carried out a one-sample proportions test, and found that workers significantly chose *QB-E* over the *QB-B*. For summaries where m=3, workers selected *QB-E* summaries 58% of the time ($p < 0.001$), and for m=4 it was of 63% ($p = 0.003$).

We also quantified the number of documents where participants preferred either the *QB-B* summary or the *QB-E* summary. We classified these assessments in three groups according to participant preferences; a *QB-B* summary was more frequently selected (*QB-B* > *QB-E*), *QB-B* and *QB-E* summaries were equally selected (*QB-B* = *QB-E*), and a *QB-E* summary was more frequently chosen (*QB-B* < *QB-E*). Results are shown in Table 6.6, and indicate that workers more frequently chose summaries using the *QB-E* method for both cases, that is, summaries containing either three or four sentences ($p < 0.001$). While results from intrinsic evaluations using a Cranfield-based approach (where relevance is a surrogate for indicativeness), and controlling for length showed that query expansion barely improved

	Sentences in common				Preferences	Proportion
	0	1	2	3		
m=3						
QB-B preferences	16	162	389	- -	567	42%
QB-E preferences	32	282	463	- -	777	58%
Total	48	444	852	- -	1,344	100%
m=4						
QB-B preferences	- -	8	13	30	51	37%
QB-E preferences	- -	4	35	48	87	63%
Total	- -	12	48	78	138	100%

Table 6.5: Number of workers who selected either QB-B or QB-E, and corresponding proportions. m represents the number of sentences that make up a summary.

Number of sentences	*QB-B* > *QB-E*	*QB-B* = *QB-E*	*QB-B* < *QB-E*	Documents
$m = 3$	73 (30%)	34 (15%)	117 (52%)	224
$m = 4$	5 (21%)	2 (9%)	16 (7%)	23

Table 6.6: Preferences per document summary.

the selection of sentences, these findings demonstrated the opposite. Users found *QB-E* summaries more useful to detect relevant documents when the length bias was removed from summaries of short and medium length summary sentences. That is, query expansion helps to compile an indicative set of sentences to generate a short query-biased summary.

We manually inspected feedback provided by participants not only to discard untrusted workers, but also to understand their opinions. We classified the feedback provided by participants into several types as shown in Table 6.7. In general, workers expressed that *QB-E* summaries included more details and focused on requests, which may have influenced their selection. Note that it could be the case that participants mentioned more than one summary feature when justifying their selection. Figure 6.2 outlines the frequency with which each category occurred. We observed that the two most popular categories were Informative and Focused, which suggested that the expansion help to locate useful sentences to construct a query-biased summary. However, sometimes participants preferred *QB-B* summaries, since they did not go off the information request. Participants were able to detect the topic drift in some summaries, a typical yet expected shortcoming of query expansion techniques. For example, observe in the figure that the category Matching request was more popular for participants selecting *QB-B* summaries.

Category	Description
Informative	Summaries were perceived informative as they conveyed more details, facts or precise information. For example, participants' opinions were: "*This summary goes into detail on topic Y*" or "*This summary gives more information on topic Y*".
Focused	The summary exhaustively covered a topic. "*This summary stayed more on topic; which helped me to better understand what was going on and in turn made me want to read the rest of the document*" "*I chose summary X because it seemed to outline the topic more closely*".
Supportive	The summary provided information additional to the topic such as causes, explanations, reasons, consequences, background or context information. For instance, "*Background information is helpful; for a subject not generally known about*" or "*Summary X explained more the situation in context*".
Coherent and Readable	The summary presented information clearly, non-fragmented or easy to understand. Typical responses regarding this feature included: "*Seems a little more fluid; without chopping up fact detail*", "*The other summary just looks like they copied different sentences off the page*". This could also mean that the summary was easy to read for workers. For example, they stated: "*This summary makes more sense to the average person reading it, seemed better written*".
Relevant	The summary discussed important information according to a topic. For instance, participants commented: "*This summary captured relevant info*" or "*I think summary X gives opinions on both sides for Topic Y. I think it has more information relevant to the question*".

Category	Description
Interesting	The summary intrigued participants to read the underlying document. Some recorded opinions were: "*Summary X presents topic Y as a controversial idea and generates more interest in the document*" or "*Seems like it would be more interesting to read*".
Matching request	The summary had a high resemblance with the information request, as some subjects mentioned: "*Summary X tells what the punishment was. This is the information I was looking for as I read the subject*" or "*Summary X more closely matches the request*".
General	The summary broadly covered information of a document. For example, participants expressed: "*This summary is better because it puts everything in perspective*" or "*Goes into more general sentences appropriate for a summary*".
Concise	The summary presented the information tightly described. Participants mentioned that "*Summary X is shorter; to the point*" or that "*Summary X is organized, clear and concise*".
Other	The summary covered other features such as technical details or appeared diverse and organised to participants. For instance: "*Summary X is a bit more diverse* " or "*Summary X seems to be better organized*".

Table 6.7: Categories of feedback given to QB-E summaries. Since participants had the choice of selecting either "Summary A" or "Summary B", we generalise the feedback in this table as "Summary X" to denote that this is the QB-E summary.

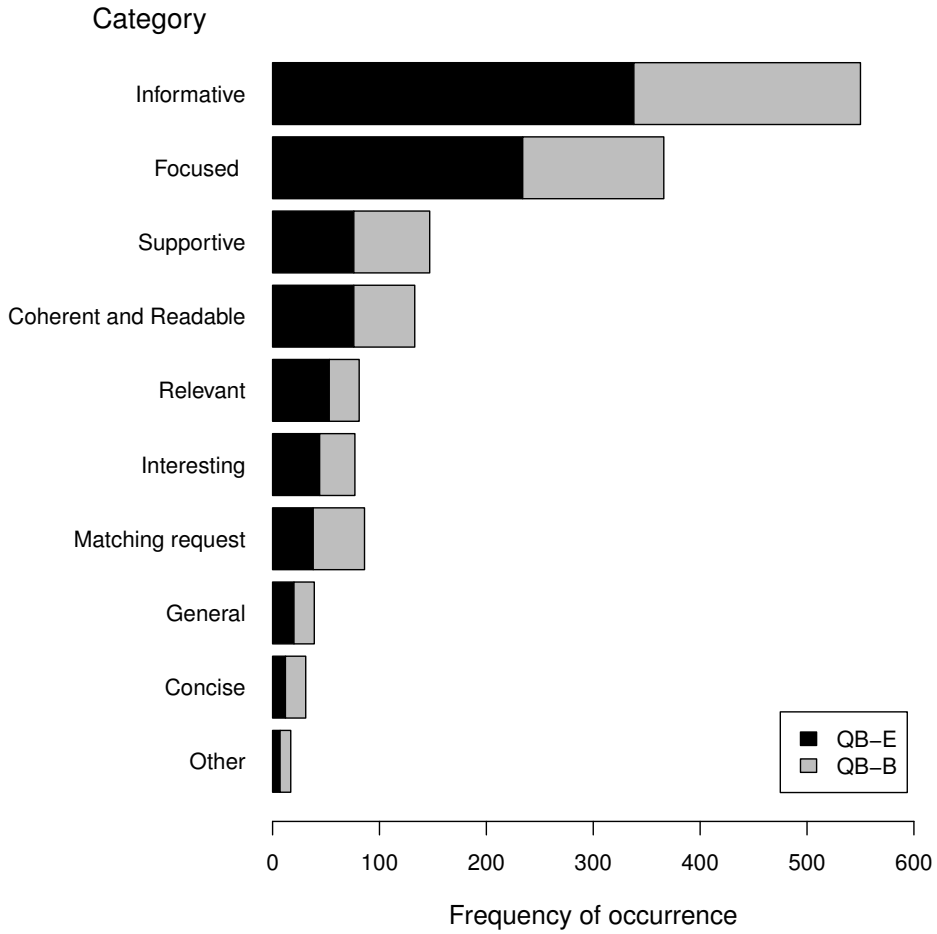

Figure 6.2: Frequency of feedback provided by participants when selecting QB-B and QB-E summaries.

6.4 Results of *QB-E* vs *Len*

In Section 5.2.4 of Chapter 5, we found that the *Len* method, which exclusively chose long sentences and ignored any other source of evidence, significantly improved the selection of relevant short sentences against a sentence ranking method that employ expansion techniques. This finding was reported in using the Novelty track 2003 dataset. We now investigate through an extrinsic summary evaluation whether the *Len* method would be a more indicative approach to construct summaries than applying query expansion to sentence ranking methods, our second hypothesis of this chapter.

There has been a series of studies exploring the perception of assessors to provide a relevance judgement when the amount of presented information is progressively increased [Barry,

1998; Janes, 1991; Marcus et al., 1978]. Other studies have found that participants prefer long answers [Kaisser et al., 2008] or long snippets [Cutrell and Guan, 2007] to assist them while carrying out informational requests. In this study, we did not aim to confirm whether subjects are biased or not towards the amount of information that is compiled in a summary. Rather, we compared if participants find summaries that employ query expansion helpful, as a mechanism to reduce the vocabulary gap problem, compared against the *Len* method. The experimental setting was described in Section 6.2, we explain our findings in this section.

A total of 75 trusted workers assessed summaries comprised of three medium sentences, and 11 trusted workers judged summaries of four short sentences. Table 6.8 shows the preferences for each ranking method, *QB-E* and *Len*. Our results indicated that query expansion was beneficial for summaries of medium length, since participants significantly selected *QB-E* summaries approximately 60% of the time (one-sample proportion test, $p < 0.001$). This finding agrees with that reported by Barry [1998]. That is, subjects are not necessarily driven by the amount of information with which are presented, rather by the content they assess. We observed that this finding also held for participants choosing *QB-E* summaries over *QB-B* summaries 60% of the time. In order to confirm that this result was not due to having same *QB-B* and *Len* summaries, on closer examination we found that 11 documents compiled the same set of sentence for both summaries. We removed assessments provided by users for these summaries, and discovered that the trend of rate selection was the same (60%).

In the case of summaries having short sentences, participants selected *QB-E* summaries 46% of the time. However, this difference was not significant (one-sample proportion test, $p = 0.359$) compared to the *Len* approach. These outcomes suggest that short sentences are less likely to convey indicative content of a document, and that query expansion will not be beneficial for this type of sentences.

Similar to the previous section, we analysed the qualitative feedback provided by participants for both summaries. These results are shown in Figure 6.3. We observed that the categories Informative, Focused, Supportive, and Coherent and Readable remained as the most frequent reasons to choose *QB-E* summaries, as in the previous analysis. However, in this study the categories Focused, Coherent and Readable, Matching request, and Concise increased their frequency of selection of *QB-E* summaries. For comparison purposes, see Figure 6.2 for feedback provided to *QB-B* and *QB-E* summaries.

		Sentences in common				Preferences	Proportion
		0	1	2	3		
m=3							
Len	preferences	80	238	222	- -	540	40%
QB-E	preferences	166	368	270	- -	804	60%
Total		246	606	492		1344	100%
m=4							
Len	preferences	3	30	28	16	75	54%
QB-E	preferences	3	30	14	14	63	46%
Total		6	60	42	30	138	100%

Table 6.8: Number of workers who selected either Len or QB-E, and corresponding proportions, where a summary is comprised of m sentences.

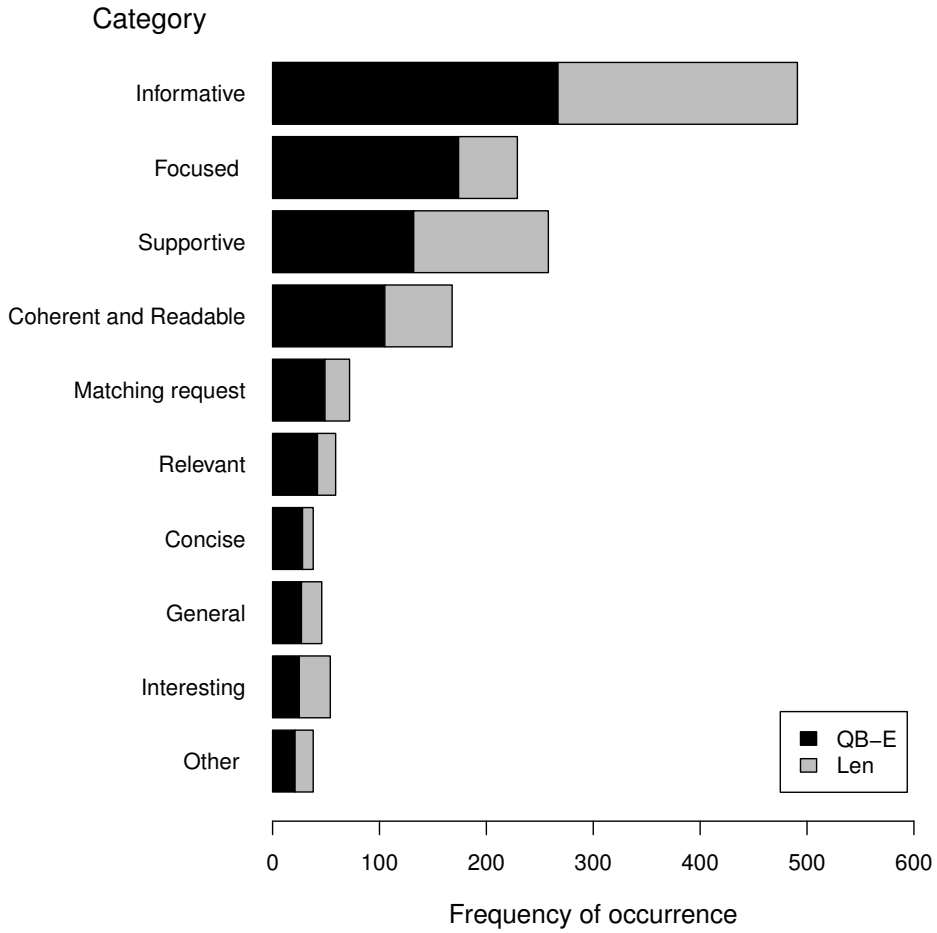

Figure 6.3: Frequency of feedback provided by participants when selecting Len and QB-E summaries.

Readability Score	m=3		m=4	
	Len	*QB-E*	*Len*	*QB-E*
Kincaid	9.6	9.4	7.2	6.9
Flesch	60.7	58.3	63.9	62.5
Fog	12.5	12.2	10.7	10.2

Table 6.9: Mean of three readability scores for Len and QB-E summaries, where m is the number of sentences in the summary.

In particular, participants expressed that *Len* summaries were sometimes difficult to read and to follow. For example, they stated "*The summary is jumpy*" or "*Summary X has extremely difficult sentence structure*". In order to investigate readability problems of *QB-E* and *Len* summaries, we computed the Kincaid index, Flesch index and Fog index, which were explained in Section 3.2.1. Table 6.9 shows the average readability scores for summaries comprising three and four sentences (m=3, 4). For summaries of three sentences, only the Flesch index indicated significant differences (t-test, $p = 0.048$). However, this was not the case for summaries of four sentences, where we did not find significant differences. From these results, we conclude that long sentences can include a wide range of information to support a request. Nevertheless, these sentences may not have a focus on the request, or may have disjoint content. For these possible reasons, participants perceived that the selected sentences were not related to make up a query-biased summary. We discuss future work regarding this observation in the next chapter.

6.5 Discussion

In this chapter, we investigated whether subjects found summaries created using query expansion techniques useful for informational requests. We mentioned at the beginning of this chapter that we grouped sentences of similar length to construct summaries. Consequently, summaries may have the appearance of compiling the same amount of information. We measured sentence length in words; however, the amount of characters in sentences may fluctuate as word length also changes. For instance, two sentences containing 17 words are less likely to contain exactly the same number of characters. Despite of grouping sentences of similar size, some of our summaries could appear longer given the word-length variability. We computed the length of summaries based on the number of characters, and found statistical differences between *QB-B* and *QB-E* (paired t-test, $p < 0.001$ for m=3, and $p = 0.004$ for m=4). We detail the character length of summaries for both sentence ranking approaches in Table 6.10.

	Mean length			Standard Deviation		
Number of sentences	*QB-B*	*QB-E*	*Len*	*QB-B*	*QB-E*	*Len*
$m = 3$	324	332	355	33	31	29
$m = 4$	225	246	305	58	60	31

Table 6.10: Mean length and standard deviation of summaries measured in characters.

	Mean length			Standard Deviation		
Number of sentences	*QB-B*	*QB-E*	*Len*	*QB-B*	*QB-E*	*Len*
$m = 3$	53	54	59	4	4	3
$m = 4$	37	39	49	9	9	5

Table 6.11: Mean length and standard deviation of summaries measured in words.

We measured the global summary size based on the number of words in sentences, see Table 6.11. The word difference between *QB-B* and *QB-E* was significant (t-test, $p = 0.002$ for m=3, and $p = 0.003$ for m=4). However, we argue that participants were unlikely to have noticed this difference: for summaries comprising 3 sentences the mean difference was of 1 word, and for summaries having 4 sentences this was of 2 words.

We also observed that both the character and word length difference was significant (t-test $p < 0.001$) between *QB-E* and *Len* summaries. These can be noted in the fourth column of Tables 6.10 and 6.11. We argue that the character and word difference is more likely to be noticed by participants. While the length could affect participants' perception to select the *Len* summary again, this was not clear in the qualitative feedback provided. Generally, participants mentioned that a summary contained more details or information. They did not explicitly express that they drove their choices because a summary appeared longer.

Our study has the limitation that participants assessed summaries only of relevant documents due to procedures of how assessors selected documents in the Novelty track. Assessors of the Novelty track firstly had to identify potentially useful documents for a given topic. They then provided relevance assessments of sentences within these documents. Thus, we assumed that documents employed in our experiment were relevant.

6.6 Summary

Query expansion is typically employed to boost the identification of possible relevant documents. In the previous chapter, we found that query expansion was useful exclusively for

short sentences based on a Cranfield evaluation approach. Conducting an extrinsic evaluation using the same testbed (Novelty track 2003 and 2004 datasets), we discovered that participants preferred summaries assisted by query expansion approximately 60% of the time, against summaries that did not employ an expanded query. These outcomes are applicable for short (5-13 words) and medium (14-20 words) length sentences.

We also found that for summaries comprised of medium length sentences, participants significantly agreed by selecting a summary that used an expanded query 60% of the time rather than a summary that merely had long sentences. For short sentences there were not significant differences between employing the expansion or selecting long sentences for this group. In particular, participants expressed that summaries simply containing long sentences method were difficult to understand, as this approach can produce a summary with disjoint ideas. In other words, long sentences does not necessarily have any relation with the user request or assist reducing the vocabulary gap problem. Based on experiments using the Novelty track dataset, we demonstrate that query expansion can benefit the selection of sentences given a specific sentence length. These sentences can make up a query-biased summary that helps users while looking for specific information.

Chapter 7

Conclusions

Summarisation is a process that aims to distill key information from a source. Based on summaries, we make decisions or form a judgement about the content of sources of information. For example, people employ short query-biased summaries displayed by search engines to identify those documents that are likely relevant to their requests, or to simply ignore those that appear irrelevant. In this thesis, we have focused on the generation of query-biased summaries for informational requests based on a sentence ranking approach, and methodologies to evaluate their effectiveness.

7.1 Research Questions and Contributions

We present our key contributions in relation to the research questions investigated in this thesis.

- *How are query-biased summaries created by humans? How does this compare with current automatic summarisation methods?*

We conducted a user study to understand how humans create short query-biased summaries by composing prose (a generative summary) and by selecting pieces from the text (an extractive summary). We introduced eye tracking techniques to track the parts of a text that were read by participants and used to write their generative summaries. In general, we observed that participants created generative summaries by generalising concepts or by combining two sentences to produce a single sentence. These types of text transformations triggered participants to employ different vocabulary that did not occur within the source text. As extractive summaries demanded less effort, we noted that generally participants followed a

top-to-bottom approach to scan the source document and to select text that became part of their summaries. To quantify the content overlap of generative and extractive summaries, we proposed to use the position of terms in both types of summaries obtained from eye tracking data. By computing the proportion of a generative summary that was covered in an extractive summary, we found that participants used similar content in both summaries around 73% of the time.

We contrasted the user study with three automatic methods to rank sentences: a cluster of significant words approach (*CL*); a query term approach (*QB*); and a combination approach that employs the former two methods and includes a position bias to score sentences (*COM*). Comparing these methods against generative summaries and using positional data, we demonstrated that these methods performed poorly. The overall coverage between human and automatic approaches was only around 22%. This reflects that current automatic methods cannot fully cover content in a document related to a given short information request.

We then evaluated summaries by ignoring positional data and measuring the coverage using a bag-of-words approach, that is, vocabulary overlap among summaries. We firstly computed the coverage between generative and extractive summaries, and found that participants' summaries content agreed around 48% of the time. This low coverage in part is due to participants referring to the same information, but using different vocabulary in their generative summaries. In contrast, automatic summaries slightly improved, showing a coverage of 24% compared with extractive summaries. Our results indicated that performance was underestimated using eye tracking data. However, it enabled us to guide our evaluation more accurately and to detect document parts that were ignored in automatic methods.

A possible criticism of our exploratory experiment is that it was constrained to a laboratory setting, with a small number of documents and queries. These limitations led us to investigate sentence ranking algorithms in a large scale testbed in our next research question.

- *How to create effective query-biased summaries?*

We continued studying the *CL*, *QB* and *COM* summarisation methods mentioned above, and included a Vector Space Model adapted for ranking sentences (*VSM*). We firstly investigated which sources of evidence were more useful to rank sentences towards a request. Using the TREC Novelty track 2003 and 2004 datasets, we discovered that methods that solely favored the selection of sentences containing query terms were more effective, that is, the *QB* and *VSM* methods. The *COM* method employed a linear combination of three sources of

evidences, so intuitively this could result in a more appealing ranking of sentences. However, when we optimised the weights for each component of the *COM* method, no statistical differences against *QB* and *VSM* were found. We also discovered that the *CL* method performed significantly more poorly compared to the other methods. For these reasons, we focused on the *QB* and *VSM* methods in subsequent experiments.

From our eye tracking experiment, we proposed to use query expansion techniques to alleviate the vocabulary mismatch problem between document content and requests of users. A query term can occur multiple times in a document; however, it is less likely that it frequently appears in a sentence. Query expansion techniques have extensively been studied for document retrieval as a mechanism to widen the pool of possible relevant results for users. We studied the relevance feedback approach proposed by Rocchio [1971] and Local Context Analysis proposed by Xu and Croft [1996; 2000]. These query expansion techniques were explored at two levels: the document level, where expanded terms were obtained from top ranked documents as typically is done for document retrieval; and at the sentence level, where the terms were sourced from top ranked sentences in highly ranked documents. Our findings indicated that the document-based expansion significantly improved the selection of relevant sentences. Depending on the ranking and query expansion approach, the increase varied from 11% to 24% using the Novelty track 2003, and from 10% to 15% in the Novelty track 2004. Significant differences between employing either Rocchio or LCA expansion techniques were not found.

We also observed that sentence-based query expansion methods significantly improved the selection of relevant sentences. However, this approach did not outperform expansion at the document level. Around 60% of the expansion terms selected using a sentence-based approach were the same as those found by employing document-based expansion. Consequently, we recommend document-based expansion approaches for sentence ranking, as they do not add complexity by requiring an extra step to rank sentences in top documents.

- *How should one evaluate sentence ranking methods using sentence-level relevance data?*

In our previous research question, we evaluated sentence ranking methods based on sentence-level relevance data. However, the TREC Novelty track was not originally created to assess summarisation tasks. We firstly investigated sentence indicativeness of the TREC Novelty track data, that is, whether a sentence conveys an indication that the content of a document is useful to further reading. By conducting a crowdsourcing study, we found that participants selected sentences judged as relevant by a TREC Novelty track assessor as being indicative

around 73% of the time. This proportion of agreement did not indicate that assessors were wrong by not agreeing 100% with TREC Novelty track relevance assessments, but rather that it is important to consider judgements in the context of specific tasks for which they have been made. We conducted an analysis to investigate the reasons for differences in relevance and indicativeness assessments. Key factors that influenced participants' assessments were: the co-reference resolution problem and lack of context in relevant sentences. Participants did not select relevant sentences that contained pronouns as indicative, since the reference was mentioned in previous sentences. We suggest that the absence of a specific reference could provide an unrelated idea for a summary. Moreover, participants valued sentences that provide supporting information or context to a request. While sentences containing this type of information were labelled as irrelevant by TREC Novelty track assessors, participants considered them as indicative for an information request.

Given that only 73% of the time sentences judged as relevant were indicative, we explored how to use this proportion of selection (indicativeness) as an assessor error rate (α). A stochastic simulation based on α indicated that the effectiveness of ranking methods decreased given sentence indicativeness compared to sentence topical relevance. However, we observed that query expansion techniques significantly improved sentence ranking methods on an indicativeness sentence basis.

While assembling summaries with the top ranked sentences for our user study, we noted that ranking methods assisted by query expansion tended to select longer sentences. Thus, a summary created from an unexpanded query was shorter on average. On closer examination of TREC Novelty track judgements, we found that assessors chose longer sentences as being relevant more frequently. In our previous research question, we ignored this length bias in the assessments during the evaluation. By introducing a length component in sentence ranking methods, which exclusively chose long sentences, we demonstrated that these methods achieved similar performance to applying query expansion techniques. This phenomenon in the TREC Novelty track data is a "win" for ranking methods that favoured long sentences, as it cloaked the main aim of query expansion techniques: to minimise the vocabulary gap problem.

We proposed to assess sentence ranking methods by explicitly controlling the length of sentences in the test collection. The evaluation consisted of grouping sentences of similar length (measured in words) therefore creating subcollections. Our results indicated that query expansion improved the ranking of short sentences (5-13 words) by around 10%. However, the expansion effect was not significant for medium (14-20 words) and long sentences

(21-29 words).

To contrast our intrinsic sentence-based evaluation, we also conducted an extrinsic evaluation to assess the selection of sentences as a unique block of information — that is, as a summary — rather than individual sentences. In this experiment, we controlled the length of sentences. In a crowdsourcing study, we investigated two questions: whether participants preferred summaries assisted by query expansion against a summary using a baseline query; and whether participants preferred summaries assisted by query expansion against a summary that merely contained long sentences. We discovered that participants significantly chose summaries assisted by query expansion 60% of the time against summaries that employed an unexpanded query, in particular for summaries compiled of medium and short sentences. We also found that participants preferred summaries using query expansion than simply long sentences 60% of the time. This finding was only applicable for medium sentences. In the case of short sentences, participants chose summaries assisted by query expansion 46% of the time. However, this difference was not significant. Our results based on these user studies confirmed that query expansion techniques assist reducing the vocabulary gap problem, since feedback provided by participants indicates that summaries that employ query expansion techniques were more informative and focused on the information request.

7.2 Future Work

In this section, we discuss possible areas for further investigation.

Eye Tracking Techniques for Studying Summarisation. In Chapter 3, we employed eye tracking techniques for analysing generative summaries. Research on reading patterns indicates that people fixating for longer periods of time may be experiencing problems in understanding a piece of information [Rayner, 1998; Reichle et al., 2006]. Therefore, a possible extension to our work is to analyse whether time is a useful indicator to prioritize areas of documents when generating summaries. For example, spending a long time in the same area may not entail that a particular area in a document is relevant to a request, but rather that is difficult to comprehend.

Query Expansion. In Chapter 4, we found that query expansion benefited the selection of relevant sentences, when these were assessed based on topical relevance. In our experiments we treated original query terms and expanded terms with the same weights. We suggest as

future work that sentence ranking methods give more priority to original terms, and gradually reduce the importance of extra terms. Recall that extra terms have certain relation with the original issued query; however, these may diverge the selection of sentences. So restricting the weight of extra terms could provide a set of sentences for assembling a query-biased summary which does not drift topically from the original request.

In our work, we did not explore efficiency aspects associated with the process of conducting query expansion or other expansion techniques. However, query expansion techniques can impose substantial overheads on a retrieval system, since queries must be evaluated twice. We leave this for future research.

Sentence Length. In Chapters 5 and 6, we controlled the length of sentences by the number of words to evaluate the effectiveness of sentence ranking methods. As an extension of this work, we suggest to measure the size of sentences by the number of "information nuggets" such as short noun-phrases or clauses instead of simple tokens. Marcus et al. [1978] investigated this approach to study if the length of bibliographic catalog fields affected the perception of users about the relevance of documents. However, this has not been investigated in terms of query-biased summaries. In Appendix E, we show that a basic normalisation approach does not clearly help to identify the sentence length bias. However, approaches such as pivoted length in a sentence context can be studied in future work.

Word Length. Former research has indicated that word-length follows a general law of language efficiency [Piantadosi et al., 2011]. That is, word-length acts better to convey informative content than word frequency. Generally, ranking methods score sentences based on word occurrence of significant terms, query terms, or cue terms, to mention a few. As further work, we advise that ranking methods can take word-length into account not only for selecting more likely informative content, but also for limiting the presentation of results.

Presentation. In Chapter 6, we created summaries by concatenating top ranked sentences, since our main aim was to select representative information towards an information request. In our user study, these summaries were presented to participants side-by-side rather than as search engines typically display results, in a ranked list. The information conveyed in several sentences may affect the typical presentation layout. Cutrell and Guan [2007] manipulated the size of snippets by presenting around 6-7 lines; however, this can vary according to the number of sentences and their length. Further work could investigate if the presentation of

summaries comprised of sentences are effective while users conduct informational requests in a large scale setting.

Readability and Cohesion. In DUC/TAC conferences the readability of summaries is gauged by assessors, see Table B.1 in Appendix B. While this approach requires human involvement, Kanungo and Orr [2009] presented a machine learning approach to predict the readability of snippets displayed by search engines. These summaries usually include truncated sentences; thus, typical readability scores may not provide a guide to assess how readable a snippet is. In Chapter 6, we computed three readability scores of summaries presented to users. Recall that in our approach the summaries comprised complete sentences. Participants expressed that summaries were difficult to read; however, readability scores did not show significant differences between summaries. We suggest that summary readability cannot fully be assessed through automatic scores that rely on sentence length or counting syllables. Rather, the cohesion among sentences can be useful to construct summaries. Cohesion is a language property that is inherent to text structure of how sentences relate to each other and connect ideas between them [Graesser et al., 2004]. Future work can employ more complex readability scores that explore linguistic features of text such as polysemy, hypernym, linking words and co-reference resolution [Graesser et al., 2004]. For example, participants in our study, detailed in Section 6.4, preferred a summary that identified the person who the summary was talking about. We leave this for future work.

7.3 Summary

In this thesis, we have investigated improvements to the sentence selection process to create short query-biased summaries (or snippets). We detect that the vocabulary mismatch problem is an obstacle for sentence ranking methods to compile an effective set of sentences from a document according to an information request. Using the TREC Novelty track as testbed, a Cranfield-based evaluation indicates that query expansion effectively assists the selection of relevant short sentences only. However, a user study shows that participants prefer summaries (comprised of either short or medium length sentences) using an expanded query 60% of the time, as these summaries provide more informative content and focus on the information request. This indicates that query expansion techniques can help in reducing the vocabulary gap, which cannot be captured through a Cranfield-based evaluation.

Our findings can serve to shape the information contained in query-biased summaries,

to provide a better understanding of the benefits of query expansion in sentence ranking approaches, and to evaluate more accurately the generation of query-biased summaries.

Appendix A

Glossary

Term	Definition
A:	Overall proportion of selection.
A_w:	Proportion of selection among workers for a given relevant sentence as indicative.
BOW:	Bag-of-words approach to evaluating summary effectiveness based on vocabulary overlapping.
CL:	Sentence ranking approach that relies on clusters of significant and non-significant terms in sentences. It uses sentence position to break ties.
COM:	Sentence ranking approach that combines three sources of evidence: clusters, query terms, and sentence position. It uses sentence position to break ties.
DC:	A document combination that represents a three sentence pairs for a given document. Each pair is comprised of one relevant and one irrelevant sentence.
$E_p^{d,q}$:	Set of words in an extractive summary by participant p for a given document-query pair d, q.
$E_s^{d,q}$:	Set of words in an automatic summary created by method s for a given document-query pair d, q.
$e_p^{d,q}$:	Set of word positions in an extractive summary by participant p for a given document-query pair d, q.
$G_p^{d,q}$:	Set of words used in a generative summary by participant p for a given document-query pair d, q.
$g_p^{d,q}$:	Set of word positions used in a generative summary by participant p for a given document-query pair d, q.

Term	Definition
l_l:	A sentence contains between 21 and 29 words.
l_m:	A sentence contains between 14 and 20 words.
l_s:	A sentence contains between 5 and 13 words.
Len:	Sentence ranking approach that is query-independent by selecting longest sentences of a document.
LCA-D:	Expansion terms using Local Context Analysis. The letter D stands for using top ranked documents to carry out the expansion process.
LCA-S:	Expansion terms using Local Context Analysis. The letter S stands for using top ranked sentences in top ranked documents to conduct the expansion process.
m:	Number of sentences returned by a sentence ranking method.
QB:	Sentence ranking approach that favours the selection of sentences containing query terms. It uses sentence position to break ties.
QB-B:	Summarisation approach that uses the *QB* sentence ranking method and unexpanded (baseline) queries.
QB-E:	Summarisation approach that uses the *QB* sentence ranking method and expanded queries (*LCA-D* approach).
QB-Len:	Sentence ranking approach that favours the selection of sentences containing query terms. It uses sentence length to break ties.
Rocchio-D:	Expansion terms using Rocchio's approach. The letter D stands for using top ranked documents to carry out the expansion process.
Rocchio-S:	Expansion terms using Rocchio's approach. The letter S stands for using top ranked sentences in top ranked documents to carry out the expansion process.
`sub2003`:	Subset of documents in the Novelty track 2003 dataset that satisfies the condition of containing at least two relevant and two non-relevant sentences in every document.
`sub2004`:	Subset of documents in the Novelty track 2004 dataset that satisfies the condition of containing at least two relevant and two non-relevant sentences in every document.
t:	Represents a topic's title of the Novelty track. It is the main baseline query.
$t+d$:	Denotes the concatenation of the title and the description field of Novelty track topics.

Term	Definition
t+n:	Denotes the concatenation of the title and the narrative field of Novelty track topics.
VSM:	Sentence ranking approach based on the Vector Space Model adapted for sentences. It uses sentence position to break ties.
VSM-Len:	Sentence ranking approach based on the Vector Space Model adapted for sentences. It uses sentence length to break ties.

Appendix B

Evaluation in DUC/TAC Conferences

Year	Collection	Summary type-task	Summary length (words)	Evaluation
2001 DUC	Newswire	Generic single document	≤100	Human and Automatic
		Generic multi-document	≤ 50,100,200 and 400	Human
		Exploratory	Non-specified	Non-specified
2002 DUC	Newswire	Generic single document abstracts	≤100	Human
		Generic multi-document abstracts	≤ 10, 50, 100 and 200	Human
		Generic multi-document extract	2 sentences (around 200-400 words)	Automatic
2003 DUC	AP, NYT and XNA	Very short single-document	≤10	Human: usefulness, and Automatic
		Short multi-document focused on events	≤100	Automatic
		Short multi-document focused on opinions	≤100	Automatic

Year	Collection	Summary type-task	Summary length (words)	Evaluation
2003 DUC	FT, FR, FBIS and LA	Short multi-document question answering	≤100	Human: responsiveness, and Automatic
2004 DUC	AP and NYT	Very short single-document	≤75 bytes	ROUGE-n gram
		Short multi-document focused on events	≤665 bytes	ROUGE-n gram
	AFP	Very short cross-lingual single-document	≤75	ROUGE-n gram
	AP and NYT	Short cross-lingual multi-document focused on events	≤665 bytes	ROUGE-n gram
	AP, NYT and XNA	Short multi-document question answering	≤665 bytes	Human: responsiveness, and Automatic
2005 DUC	FT and LA	Multi-document given a specific request and user profile	≤250	Human: responsiveness and quality, Automatic: ROUGE-1, ROUGE-2 and ROUGE-SU4
2006 DUC	AQUAINT	Multi-document given a specific request	≤250	Human: responsiveness and quality, Automatic: ROUGE-2 and ROUGE-SU4
2007 DUC	AQUAINT	Multi-document given a specific request	≤250	Human: responsiveness, quality and the Pyramid method, Automatic: ROUGE-2, ROUGE-SU4 and BE
		Multi-document updated summary	≤100	Human: the Pyramid method, Automatic: ROUGE-2, ROUGE-SU4 and BE
2008 TAC	AQUAINT-2	Multi-document given a specific request	≤100	Human: responsiveness, readability and the Pyramid method
		Multi-document updated given a specific request	≤100	

Year	Collection	Summary type-task	Summary length (words)	Evaluation
	TAC 2008 QA Questions	Single-document opinion	Non-specified	Human: responsiveness and the Pyramid method
2009 TAC	AQUAINT-2	Multi-document given a specific request Multi-document updated given a specific request	≤100 ≤100	Human: responsiveness, readability and the Pyramid method
2010 TAC	AQUAINT and AQUAINT-2	Multi-document describing an event Multi-document updated	≤100 ≤100	Human: responsiveness, readability and the Pyramid method
2011 TAC	AP(2), NYT(2) and XNA(2)	Multi-document describing an event Multi-document updated	≤100 ≤100	Human: responsiveness, readability and the Pyramid method
	WikiNews in 7 languages	Multi-document and multi-lingual	240-250	Human: responsiveness, Automatic: ROUGE

Table B.1: Table with different years of DUC/TAC summarisation tasks. The summary length is given in words, otherwise is stated. AP= Associated Press newswire (1998-2000), NYT= New York Times newswire (1998-2000), XNA= Xinhua News Agency (English version, 1996-2000), FT=Finantial Times of London (1991-1994), FR=Federal Register (1994), FBIS (1996), LA=Los Angeles Times (1989-1990), AFP=Agence France Press, AP(2)=Associated Press newswire (2007-2008), NYT(2)= New York Times newswire (2007-2008) and XNA(2)= Xinhua News Agency (English version, 2007-2008). The AQUAINT corpus includes newspapers articles from Associated Press (1998-2000), the New York Times (1998-2000) and Xinhua News Agency (English version, 1996-2000). The AQUAINT-2 corpus contains reports from Agence France Press, Central News Agency, Xinhua Agency, Los Angeles Times, Washington Post News Service, the New York Times and the Associated Press (2004-2006). In the early two years of DUC tracks the collections are unknown, as reviews only mentioned that these were of newswire nature.

Appendix C

Second Simulated Work Task and Indicative Requests

For the topics *Antarctica expeditions* and *Oceanographic vessels*, we created the following simulated work task, with indicative requests detailed in Table C.1.

> *As member of a scientific group, you will undertake an expedition to Antarctica. You have been asked to research prior expeditions and the roll of krill in Antarctica. Your team leader requires that you write a short summary of any document you come across that answers their questions.*

Original Request	Simulated Scenario
Description of Topic 353: Identify systematic explorations and scientific investigations of Antarctica, current or planned.	Indicative Request 1. Your team leader is interested in the following information: *Identify current or planned systematic explorations and scientific investigations of Antarctica.*
Description of Topic 399: Identify documents that discuss the activities or equipment of oceanographic vessels.	Indicative Request 2. Your team leader is interested in the following information: *What are krill and why are they important to Antarctica?*

Table C.1: Indicative requests based on two TREC queries. Since the chosen document includes information about krill, we modified the indicative requests according to the document content.

Appendix D

Survey

Survey conducted to gather demographic information and general search engine use from volunteers that participated in our user study, which was detailed in Chapter 3. Numbers indicate the frequency of occurrence for each question from a total of 10 participants.

Required*

Age*

18-22	2
23-28	6
29-33	1
34-40	1
Older	0

Gender*

Female	0
Male	10

Educational level*

Undergraduate	5
Graduate	1
Postgraduate	4
Other	0

Area of study*

Computer Science or IT	9
Business	1

Languages spoken*

Chinese	4
English	10
Italian	1
Hindi	2
Serbian	1

Which search engine do you use more often?*

Google	10
Yahoo!	0
Bing	0
Other	1

How useful are Web search engine summaries to guied your search?*

Extremely not useful	0
Not useful	0
Neutral	1
Useful	3
Extremely useful	6

How often do you carry out online searches?*

Less than once a week	0
Once or twice a week	0
Once or twice a day	0
More than once or twice a day	10
Other	0

Appendix E

Normalisation

Normalisation is an approach used to minimise the effect of favouring the retrieval of long documents [Singhal et al., 1996]. In this appendix, we outline results after applying a basic normalisation technique to sentence ranking methods. Sentence ranking approaches may employ certain mechanisms to reduce the sentence length bias. For example, the *QB* scoring formula in Equation 2.12 (see page 86) uses the total number of words in a query (tq), which can be considered as a normalisation factor. However, the scoring formula can not completely isolate the effect of favoring long sentences to achieve higher P@2 values.

We normalised sentence scores for ranking methods that resolved ties by the position of a sentence (*QB* and *VSM*), and methods that employed the length of a sentence (*QB-Len* and *VSM-Len*). The normalisation consisted of dividing a sentence score given by a ranking method by the total number of words in that sentence [Tsegay et al., 2009]. Table E.1 summarises averaged P@2 over topics of normalised sentence scores. From these results, we made the following observations:

- Basic normalisation significantly decreased performance in comparison to non-normalised sentence scores (paired Wilcoxon test $p < 0.001$). This trend was noted in *QB*, *VSM*, *QB-Len* or *VSM-Len* using any of the baselines (t,t+d, t+n) as well as using *Rocchio-D* or *LCA-D* expansion terms. For direct comparison, the performance of non-normalised sentence scores has been reported in Table 4.4 for *QB* and *VSM*. Results for *QB-Len* and *VSM-Len* without expansion were detailed in Tables 5.9, and employing query expansion in Table 5.11.

- Basic normalisation does not clearly help to identify the length bias in sentence ranking. In Table E.1, the symbol ↑ denotes that the expansion approach obtained a positive percentage change with respect to the title (t) of a topic as the baseline query. In contrast, the symbol ↓ represents that the expansion performed under the baseline query. We noted that the LCA document-based technique significantly improved sentence ranking approaches (*QB*, *VSM*, *QB-Len* and *VSM-Len*) for the `sub2003` and `sub2004` datasets. Significance values are detailed in Table E.1. The enhancement ranges between 12% and 38% when ranking methods

Query	***QB***		***VSM***		***QB-Len***		***VSM-Len***	
	sub2003	sub2004	sub2003	sub2004	sub2003	sub2004	sub2003	sub2004
t	0.44	0.33	0.50	0.33	0.45	0.33	0.51	0.33
t+d	0.42	0.35	0.45	0.34	0.43	0.35	0.45	0.34
t+n	0.49	0.39	0.52	0.40	0.50	0.39	0.53	0.40
t+Rocchio-D	0.48 ↑	0.38† ↑	0.40† ↓	0.31 ↓	0.49 ↑	0.38† ↑	0.40‡ ↓	0.31 ↓
t+LCA-D	0.56‡↑	0.45‡ ↑	0.57‡ ↑	0.44‡ ↑	0.57‡ ↑	0.45‡ ↑	0.57† ↑	0.44‡ ↑

Table E.1: Averaged P@2 over topics of normalised sentence scores using four different sentence ranking methods: QB, VSM, QB-Len and VSM-Len. The symbol ↑ shows an increase against the baseline query title (t), and the symbol ↓ denotes a decrease against the same baseline. Paired Wilcoxon test at significance values of 0.05 (†) and 0.001 (‡).

employed the expansion against the title baseline. These results are consistent with those reported in Section 4.4.2, where sentence ranking approaches ignore the length bias. We also note that using *LCA-D* terms for expansion, the percentage increase was significant with respect to the title baseline (see fifth row in Table E.1). However, these findings contradicted results described in Section 5.2.3, where the *LCA-D* expansion only improved around 5% in the `sub2004` dataset.

- After normalising sentence scores, there was no significant difference in performance between *QB* and *QB-Len*, or *VSM* and *VSM-Len* with their corresponding baseline or query expansion counterpart. This occurred for all query baselines and expansion approaches.

Bibliography

S. Afantenos, V. Karkaletsis, and P. Stamatopoulos. Summarization from medical documents: a survey. *Artificial Intelligence in Medicine*, 33(2):157–177, 2005.

H. Al-Maqbali, F. Scholer, J. A. Thom, and M. Wu. Evaluating the effectiveness of visual summaries for web search. In *Proceedings of the 15th Australasian Document Computing Symposium*, pages 36–43, 2010.

J. Allan, C. Wade, and A. Bolivar. Retrieval and novelty detection at the sentence level. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 314–321, New York, NY, USA, 2003. ACM.

O. Alonso and S. Mizzaro. Can we get rid of TREC assessors? Using Mechanical Turk for relevance assessment. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*, pages 15–16, 2009.

M. R. Amini and N. Usunier. A contextual query expansion approach by term clustering for robust text summarization. In *7th Document Understanding Conference (DUC'07)*. PASCAL - Pattern Analysis, Statistical Modelling and Computational Learning, 2007.

M. R. Amini, N. Usunier, and P. Gallinari. Automatic text summarization based on word-clusters and ranking algorithms. In *27th European Conference on IR Research*, pages 142–156. Springer Berlin Heidelberg, 2005.

E. Amitay and C. Paris. Automatically summarising Web sites: is there a way around it? In *Proceedings of the ninth international conference on Information and knowledge management*, pages 173–179, New York, NY, USA, 2000. ACM.

R. Baeza-Yates and B. Ribeiro-Neto. *Modern information retrieval*. ACM/Addison Wesley, 1999.

N. Balasubramanian, J. Allan, and W. B. Croft. A comparison of sentence retrieval techniques. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 813–814, New York, NY, USA, 2007. ACM.

L. L. Bando, F. Scholer, and A. Turpin. Constructing query-biased summaries: a comparison of human and system generated snippets. In *Proceedings of the third symposium on Information interaction in context*, pages 195–204, New York, NY, USA, 2010. ACM.

L. L. Bando, F. Scholer, and A. Turpin. Sentence Length Bias in TREC Novelty Track Judgements. In *Proceedings of the 17th Australasian Document Computing Symposium 2012*, pages 55–61. ACM, 2012.

C. L. Barry. Document representations and clues to document relevance. *Journal of the American Society for Information Science*, 49(14):1293–1303, 1998.

R. Barzilay and M. Elhadad. Using lexical chains for text summarization. In *Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization*, volume 17, pages 10–17, 1997.

P. B. Baxendale. Machine-made index for technical literature –an experiment. *IBM Journal of Research and Development*, 2(4):354–361, 1958.

M. Bendersky and W. B. Croft. Analysis of long queries in a large scale search log. In *Proceedings of the 2009 workshop on Web Search Click Data*, pages 8–14, New York, NY, USA, 2009. ACM.

A. Berger and V. O. Mittal. Query-relevant summarization using FAQs. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 294–301, Stroudsburg, PA, USA, 2000. ACL.

S. Berkovsky, T. Baldwin, and I. Zukerman. Aspect-based personalized text summarization. In *Adaptive Hypermedia and Adaptive Web-Based Systems*, pages 267–270. Springer Berlin Heidelberg, 2008.

J. Bhogal, A. Macfarlane, and P. Smith. A review of ontology based query expansion. *Information Processing & Management*, 43:866–886, 2007.

B. Billerbeck. *Efficient Query Expansion*. PhD thesis, RMIT University, 2005.

B. Billerbeck and J. Zobel. When query expansion fails. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 387–388, New York, NY, USA, 2003. ACM.

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

P. Borlund. The concept of relevance in IR. *Journal of the American Society for Information Science and Technology*, 54(10):913–925, 2003.

R. Braddock. The frequency and placement of topic sentences in expository prose. *Research in the Teaching of English*, 8(3):287–302, 1974.

R. Brandow, K. Mitze, and L. F. Rau. Automatic condensation of electronic publications by sentence selection. *Information Processing & Management*, 31(5):675–685, 1995.

A. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002.

A. L. Brown and J. D. Day. Macrorules for summarizing texts: The development of expertise. *Journal of Verbal Learning and Verbal Behavior*, 22:1–14, 1983.

P. Buchheit. Variable length snippet generation. U.S. Patent Application 10/866,466, 2005.

C. Buckley. Why current IR engines fail. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 584–585, New York, NY, USA, 2004. ACM.

C. Buckley, G. Salton, J. Allan, and A. Singhal. Automatic query expansion using SMART: TREC 3. In *Overview of the Third Text REtrieval Conference (TREC-3)*, pages 69–80, 1995.

G. Buscher, A. Dengel, and L. van Elst. Query expansion using gaze-based feedback on the subdocument level. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 387–394, New York, NY, USA, 2008. ACM.

S. Büttcher, C. L. A. Clarke, and G. V. Cormack. *Information Retrieval: Implementing and evaluating search engines*. MIT Press, 2010.

J. P. Callan. Passage-level evidence in document retrieval. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 302–310, New York, NY, USA, 1994. Springer-Verlag.

C. Callison-Burch. Fast, cheap, and creative: evaluating translation quality using Amazon's Mechanical Turk. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 286–295, Stroudsburg, PA, USA, 2009. ACL.

C. S. Campbell and P. P. Maglio. A robust algorithm for reading detection. In *Proceedings of the 2001 workshop on Perceptive user interfaces*, pages 1–7, New York, NY, USA, 2001. ACM.

G. Cao, J. Nie, J. Gao, and S. Robertson. Selecting good expansion terms for pseudo-relevance feedback. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 243–250, New York, NY, USA, 2008. ACM.

C. Carpineto and G. Romano. A survey of automatic query expansion in information retrieval. *ACM Computing Surveys*, 44(1):1–50, 2012.

C. Clarke, N. Craswell, and I. Soboroff. Overview of the TREC 2004 Terabyte Track. In *Proceedings of TREC 2004*, 2004.

C. L. A. Clarke, E. Agichtein, S. Dumais, and R. W. White. The influence of caption features on clickthrough patterns in web search. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 135–142, New York, NY, USA, 2007. ACM.

C. Cleverdon. The Cranfield tests on index language devices. *ASLIB Proceedings*, 19:173–194, 1967.

J. Cohen. *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates, second edition, 1988.

N. Craswell, D. Hawking, and S. Robertson. Effective site finding using link anchor information. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 250–257, New York, NY, USA, 2001. ACM.

W. B. Croft and D. J. Harper. Using probabilistic models of document retrieval without relevance. *Journal of Documentation*, 35:285–295, 1979.

W. B. Croft, D. Metzler, and T. Strohman. *Search Engines: Information retrieval in practice*. Addison Wesley, 2009.

E. Cutrell and Z. Guan. What are you looking for? an eye-tracking study of information usage in web search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 407–416, New York, NY, USA, 2007. ACM.

H. T. Dang. Overview of DUC 2005. In *Proceedings of the Document Understanding Conference*, 2005.

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

A. Díaz and P. Gervás. User-model based personalized summarization. *Information Processing & Management*, 43(6):1715–1734, 2007.

R. L. Donaway, K. W. Drummey, and L. A. Mather. A comparison of rankings produced by summarization evaluation measures. In *Proceedings of the 2000 NAACL-ANLP Workshop on Automatic summarization - Volume 4*, pages 69–78, Stroudsburg, PA, USA, 2000. ACL.

D. Donlan. Locating main ideas in history textbooks. *Journal of Reading*, 24(2):135–140, 1980.

A. Duchowski. A breadth-first survey of eye-tracking applications. *Behavior Research Methods*, 34 (4):455–470, 2002.

A. Duchowski. *Eye Tracking Methodology: Theory and Practice*. Springer, second edition, 2007.

H. P. Edmundson. New methods in automatic extracting. *Journal of the ACM*, 16(2):264–285, 1969.

H. P. Edmundson and R. E. Wyllys. Automatic abstracting and indexing - survey and recommendations. *Communications of the ACM*, 4:226–234, 1961.

E. N. Efthimiadis. Query expansion. *Annual review of information science and technology*, 31:121–187, 1996.

D. E. Egan, J. R. Remde, L. M. Gomez, T. K. Landauer, J. Eberhardt, and C. C. Lochbaum. Formative design evaluation of Superbook. *ACM Transactions on Information Systems*, 7(1):30–57, 1989.

A. El-Hamdouchi and P. Willet. Comparison of hierarchic agglomerative clustering methods for document retrieval. *The Computer Journal*, 32(3):220–227, 1989.

G. Erkan and D. R. Radev. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479, 2004.

K. N. Fachry, J. Kamps, and J. Zhang. The impact of summaries: What makes a user click? In *Proceedings of the 10th Dutch-Belgian Information Retrieval Workshop*, pages 47–54, 2010.

C. Fellbaum. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press, 1998.

M. Fiszman, T. Rindflesch, and H. Kilicoglu. Abstraction summarization for managing the biomedical research literature. In *Proceedings of the HLT-NAACL Workshop on Computational Lexical Semantics*, pages 76–83. ACL, 2004.

D. Fum, G. Guida, and C. Tasso. Forward and backward reasoning in automatic abstracting. In *Proceedings of the 9th conference on Computational linguistics - Volume 1*, pages 83–88. Academia Praha, 1982.

G. W. Furmas, T. K. Landauer, L. M. Gomez, and S. T. Dumais. The vocabulary problem in human-system communication. *Communications of the ACM*, 30(11):964–971, 1987.

R. Gaizauskas, J. Foster, Y. Wilks, J. Arundel, P. Clough, and S. Piao. The METER corpus: a corpus for analysing journalistic text reuse. In *Proceedings of the Corpus Linguistics 2001 Conference*, pages 214–223, 2001.

A. Garg and K. Dhamdhere. Session-based dynamic search snippets. U.S. Patent 8,145,630, 2012.

R. Garner. Efficient text summarization: Costs and benefits. *The Journal of Educational Research*, 75(5):275–279, 1982.

J. Goldstein, M. Kantrowitz, V. Mittal, and J. Carbonell. Summarizing text documents: sentence selection and evaluation metrics. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 121–128, New York, NY, USA, 1999. ACM.

B. Gomes and B. T. Smith. Detecting query-specific duplicate documents. U.S. Patent 6,615,209, 2003.

A. C. Graesser, D. McNamara, M. M. Louwerse, and Z. Cai. Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, 26(2):193–202, 2004.

G. Grefenstette. Producing intelligent telegraphic text reduction to provide an audio scanning service for the blind. In *Working notes of the AAAI Spring Symposium on Intelligent Text summarization*, pages 111–118, 1998.

Y. Guo, H. Harkema, and R. Gaizauskas. Sheffield University and the TREC 2004 Genomics Track: query expansion using synonymous terms. In *Proceedings of the Thirteenth Text REtrieval Conference (TREC-13)*, pages 753–757, 2004.

H. Halteren and S. Teufel. Examining the consensus between human summaries: initial experiments with factoid analysis. In *Proceedings of the HLT-NAACL 03 on Text summarization workshop*, volume 5, pages 57–64, Stroudsburg, PA, USA, 2003. ACL.

K. S. Han, D. H. Baek, and H. C. Rim. Automatic text summarization based on relevance feedback with query splitting. In *Proceedings of the fifth international workshop on on Information retrieval with Asian languages*, pages 201–202, New York, NY, USA, 2000. ACM.

T. F. Hand. A proporsal for task-based evaluation of text summarization system. In *Proceedings of the TIPSTER Text Phase III Workshop*, pages 31–38, 1997.

D. Harman. Relevance feedback revisited. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 1–10, New York, NY, USA, 1992. ACM.

D. Harman. Overview of the TREC 2002 Novelty Track. In *Proceedings of TREC 2002*, 2002.

J. A. Hartigan and M. A. Wong. Algorithm AS 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society*, 28(1):100–108, 1979.

V. Hatzivassiloglou, J. L. Klavans, M. L. Holcombe, R. Barzilay, M. Kan, and K. R. McKeown. SIMFINDER: A flexible clustering tool for summarization. In *Proceedings of the NAACL Workshop on Automatic Summarization*, pages 41–49, 2001.

M. A. Hearst. TextTiling: segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64, 1997.

W. Hersh, S. Price, and L. Donohoe. Assessing thesaurus-based query expansion using the UMLS Metathesaurus. In *Proceedings of the AMIA Symposium*, pages 344–348. American Medical Informatics Association, 2000.

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

E. Hovy and C. Y. Lin. Automated text summarization and the SUMMARIST system. In *TIPSTER*, pages 197–214. ACL, 1998.

E. Hovy, C. Y. Lin, L. Zhou, and J. Fukumoto. Automated summarization evaluation with basic elements. In *Proceedings of the Fifth Conference on Language Resources and Evaluation*, pages 604–611, 2006.

B. L. Humphreys and D. A. Lindberg. Building the unified medical language system. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, pages 475–480, 1989.

J. Hutchins. Summarization: Some problems and methods. *Meaning: The frontier of Informatics*, 9: 151–173, 1987.

J. Hyönä and A. Nurminen. Do adult readers know how they read? Evidence from eye movement patterns and verbal reports. *British Journal of Psychology*, 97:31–50, 2006.

J. W. Irwin and M. A. Doyle. *Reading/Writing Connections: Learning from Research*. International Reading Association, 1992.

J. W. Janes. Relevance judgements and the incremental presentation of document representations. *Information Processing & Management*, 27(6):629–646, 1991.

B. Jansen, D. L. Booth, and A. Spink. Determining the user intent of web search engine queries. In *Proceedings of the 16th international conference on World Wide Web*, pages 1149–1150, New York, NY, USA, 2007. ACM.

H. Jing and K. R. McKeown. The decomposition of human-written summary sentences. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 129–136, New York, NY, USA, 1999. ACM.

H. Jing and K. R. McKeown. Cut and paste based text summarization. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pages 178–185, Stroudsburg, PA, USA, 2000. ACL.

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

Y. Jing and W. B. Croft. An association thesaurus for information retrieval. In *Proceedings of RIAO*, pages 146–160, 1994.

T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 154–161, New York, NY, USA, 2005. ACM.

H. Joho, D. Hannah, and J. M. Jose. Emulating query-biased summaries using document titles. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 709–710, New York, NY, USA, 2008. ACM.

M. Kaisser, M. A. Hearst, and J. B. Lowe. Improving search results quality by customizing summary lengths. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, pages 701–709. ACL, 2008.

S. Kallurkar, Y. Shi, R. S. Cost, C. Nicholas, A. Java, C. James, S. Rajavaram, V. Shanbhag, S. Bhatkar, and D. Ogle. UMBC at TREC 12. In *The 12th Text Retrieval Conference (TREC-2003)*, 2003.

T. Kanungo and D. Orr. Predicting the readability of short web summaries. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 202–211, New York, NY, USA, 2009. ACM.

D. Kelly. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval*, 3(1-2):1–224, 2009.

E. Kintsch. Macroprocesses and microprocesses in the development of summarization skill. *Cognition and Instruction*, 7(3):161–195, 1990.

K. Knight and D. Marcu. Summarization beyond sentence extraction: A probabilistic approach to sentence compression. *Artificial Intelligence*, 139(1):91–107, 2002.

Y. Ko, H. An, and J. Seo. Pseudo-relevance feedback and statistical query expansion for web snippet generation. *Information Processing Letters*, 109(1):18–22, 2008.

R. Krovetz. Viewing morphology as an inference process. In *Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 191–202, New York, NY, USA, 1993. ACM.

J. Kupiec, J. Pedersen, and F. Chen. A trainable document summarizer. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 68–73, New York, NY, USA, 1995. ACM.

X. Li and W. B. Croft. Novelty detection based on sentence level patterns. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 744–751, New York, NY, USA, 2005. ACM.

C. Y. Lin. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization*, pages 74–81, 2004.

C. Y. Lin and E. Hovy. Identifying topics by position. In *Proceedings of the fifth conference on Applied natural language processing*, pages 283–290, Stroudsburg, PA, USA, 1997. ACL.

S. P. Liversedge and J. M. Findlay. Saccadic eye movements and cognition. *Trends in Cognitive Sciences*, 4(1):6–14, 2000.

E. Lloret and M. Palomar. Text summarisation in progress: a literature review. *Artificial Intelligence Review*, 37(1):1–41, 2012.

L. Lorigo, M. Haridasan, H. Brynjarsdóttir, L. Xia, T. Joachims, G. Gay, L. Granka, F. Pellacini, and B. Pan. Eye tracking and online search: Lessons learned and challenges ahead. *Journal of the American Society for Information Science and Technology*, 59(7):1041–1052, 2008.

D. E. Losada. Statistical query expansion for sentence retrieval and its effects on weak and strong queries. *Information Retrieval*, 13(5):485–506, 2010.

D. E. Losada, L. Azzopardi, and M. Baillie. Revisiting the relationship between document length and relevance. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 419–428, New York, NY, USA, 2008. ACM.

H. P. Luhn. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165, 1958.

I. Mani and E. Bloedorn. Machine learning of generic and user-focused summarization. In *Proceedings of the National Conference on Artificial Intelligence, (AAAI)*, pages 821–826. John Wiley & Sons LTD, 1998.

I. Mani, D. House, G. Klein, L. Hirschman, T. Firmin, and B. Sundheim. The TIPSTER SUMMAC Text Summarization Evaluation. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, pages 77–85, Stroudsburg, PA, USA, 1999. ACL.

D. Marcu. From discourse structures to text summaries. In *Proceedings of the ACL*, pages 82–88, 1997.

D. Marcu. The automatic construction of large-scale corpora for summarization research. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 137–144, New York, NY, USA, 1999. ACM.

R. S. Marcus, P. Kugel, and A. R. Benenfeld. Catalog information and text as indicators of relevance. *Journal of the American Society for Information Science*, 29(1):15–30, 1978.

K. R. McKeown, R. J. Passonneau, D. K. Elson, A. Nenkova, and J. Hirschberg. Do summaries help? A task-based evaluation of multi-document summarization. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 210–217, New York, NY, USA, 2005. ACM.

D. Metzler and T. Kanungo. Machine learned sentence selection strategies for query-biased summarization. In *SIGIR Learning to Rank Workshop*, pages 40–47, 2008.

G. A. Miller. A Lexical Database for English. *Communications of the ACM*, 38(11):39–41, 1995.

S. Mizzaro. Relevance: The Whole History. *Journal of the American Society for Information Science*, 48(9):810–832, 1997.

C. Monz. *From Document Retrieval to Question Answering*. PhD thesis, University of Amsterdam, 2003.

V. Murdock. *Aspects of sentence retrieval*. PhD thesis, University of Massachusetts Amherst, 2006.

V. Murdock and W. B. Croft. A translation model for sentence retrieval. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 684–691, Stroudsburg, PA, USA, 2005. ACL.

A. Nenkova and K. McKeown. *Automatic Summarization*. Now Publishers Inc., 2011.

A. Nenkova and R. Passonneau. Evaluationg content selection in summarization: The pyramid method. In *Proceedings of Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 145–152, 2004.

J. Nielsen. F-shaped pattern for reading web content, April 2006. URL `http://www.useit.com/alertbox/reading_pattern.html`. Accessed August 23, 2012.

C. D. Paice. Constructing literature abstracts by computer: techniques and prospects. *Information Processing & Management*, 26(1):171–186, 1990.

J. Pedersen, D. Cutting, and J. Tukey. Snippet search: A single phrase approach to text access. In *Proceedings of the 1991 Joint Statistical Meetings*, 1991.

T. Petersen. *Art & architecture thesaurus*. Oxford University Press, 1990.

S. T. Piantadosi, H. Tily, and E. Gibson. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9):3526–3529, 2011.

M. Porter. An algorithm for suffix stripping. *Program: electronic library and information systems*, 14(3):130–137, 1980.

D. R. Radev and W. Fan. Automatic summarization of search engine hit lists. In *Proceedings of the ACL-2000 workshop on Recent advances in natural language processing and information retrieval*, pages 99–109. ACM, 2000.

D. R. Radev and K. R. McKeown. Generating natural language summaries from multiple on-line sources. *Computational Linguistics*, 24:470–500, 1998.

D. R. Radev and D. Tam. Summarization evaluation using the relative utility. In *Proceedings of the twelfth international conference on Information and knowledge management*, pages 508–511, New York, NY, USA, 2003. ACM.

D. R. Radev, H. Jing, M. Styśś, and D. Tam. Centroid-based summarization of multiple documents. *Information Processing & Management*, 40(6):Elsevier Ltd., 2004.

G. J. Rath, A. Resnick, and T. R. Savage. The formation of abstracts by the selection of sentences. Part I. Sentence selection by men and machines. *American Documentation*, 12(2):139–141, 1961.

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

E. D. Reichle, A. Pollatsek, and K. Rayner. E-Z Reader: A cognitive-control, serial-attention model of eye-movement behavior during reading. *Cognitive Systems Research*, 7:4–22, 2006.

S. E. Robertson and K. Spärck-Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27:129–146, 1976.

S. E. Robertson, S. Walker, M. M. Beaulieu, M. Gatford, and A. Payne. Okapi at TREC-3. In *Overview of the Third Text Retrieval Conference (TREC-3)*, pages 109–126, 1995.

J. J. Rocchio. *Relevance feedback in information retrieval*, pages 313–323. Prentice Hall Inc., Englewood Cliffs, NJ, 1971.

D. E. Rose, D. Orr, and R. G. Prasad. Summary attributes and perceived search quality. In *Proceedings of the 16th international conference on World Wide Web*, pages 1201–1202, New York, NY, USA, 2007. ACM.

I. Ruthven. Re-examining the potential effectiveness of interactive query expansion. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 213–220, New York, NY, USA, 2003. ACM.

H. Saggion. Automatic summarization: an overview. *Revue française de linguistique appliquée*, XIII: 63–81, 2008.

G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523, 1988.

G. Salton and C. Buckley. Improving retrieval performance by relevance feedback. *Journal of the American Society for Information Science*, 41(4):288–297, 1990.

G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

G. Salton, A. Singhal, M. Mitra, and C. Buckley. Automatic text structuring and summarization. *Information Processing & Management*, 33(2):193–207, 1997.

M. Sanderson. Accurate user directed summarization from existing tools. In *Proceedings of the seventh international conference on Information and knowledge management*, pages 45–51, New York, NY, USA, 1998. ACM.

T. Saracevic. Relevance: A review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science*, 26(6):321–343, 1975.

T. Saracevic. Relevance: A Review of the Literature and a Framework for Thinking on the Notion in the Information Science. Part II: Nature and Manifestations of Relevance. *Journal of the American Society for Information Science and Technology*, 58(13):1915–1933, 2007a.

T. Saracevic. Relevance: A Review of the Literature and a Framework for Thinking on the Notion in the Information Science. Part III: Behaviour and Effects of Relevance. *Journal of the American Society for Information Science and Technology*, 58(13):2126–2144, 2007b.

F. Scholer, H. E. Williams, and A. Turpin. Query association surrogates for web search. *Journal of the American Society for Information Science and Technology*, 55(7):637–650, 2004.

F. Scholer, A. Turpin, and M. Sanderson. Quantifying test collection quality based on the consistency of relevance judgements. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 1063–1072, New York, NY, USA, 2011. ACM.

A. Singhal, G. Salton, M. Mitra, and C. Buckley. Document length normalization. *Information Processing & Management*, 32:619–633, 1996.

A. F. Smeaton and C. J. van Rijsbergen. The retrieval effects of query expansion on a feedback document retrieval system. *The Computer Journal*, 26(3):239–246, 1983.

M. D. Smucker and J. Allan. An investigation of Dirichlet prior smoothings performance advantage. Technical report, The University of Massachusetts, The Center for Intelligent Information Retrieval, 2005.

R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Stroudsburg, PA, USA, 2008. ACL.

I. Soboroff. Overview of the TREC 2004 Novelty Track. In *The Thirteenth Text Retrieval Conference (TREC 2004)*, 2004.

I. Soboroff and D. Harman. Overview of the TREC 2003 Novelty Track. In *Proceedings of the Twelfth Text Retrieval Conference (TREC 2003)*, 2003.

K. Spärck-Jones. *Automatic keyword classification for information retrieval*. Butterworths, 1971.

K. Spärck-Jones. Automatic summarising: The state of the art. *Information Processing & Management*, 43(6):1449–1481, 2007.

K. Spärck-Jones, S. Walker, and S. E. Robertson. A probabilistic model of information retrieval: development and comparative experiments. *Information Processing & Management*, 36(6):779–808, 2000.

C. Sporleder and M. Lapata. Discourse chunking and its application to sentence compression. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 257–264, Stroudsburg, PA, USA, 2005. ACL.

T. Strzalkowski, J. Wang, and W. B. Summarization-based query expansion in information retrieval. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 2*, pages 1258–1264, Stroudsburg, PA, USA, 1998. ACL.

F. M. Suchanek, G. Kasneci, and G. Weikum. YAGO: A Large Ontology from Wikipedia and WordNet. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(3):203–217, 2008.

J. Sun, D. Shen, H. Zeng, Q. Yang, Y. Lu, and Z. Chen. Web-page summarization using clickthrough data. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 194–201, New York, NY, USA, 2005. ACM.

J. Teevan, E. Cutrell, D. Fisher, S. M. Drucker, G. Ramos, P. André, and C. Hu. Visual snippets: summarizing web pages for search and revisitation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2023–2032, New York, NY, USA, 2009. ACM.

S. Teufel and M. Moens. Sentence extraction as a classification task. In *Proceedings of the ACL*, pages 58–65, 1997.

K. W. Thiede and M. C. Anderson. Summarizing can improve metacomprehension accuracy. *Contemporary Educational Psychology*, 28(2):129–160, 2003.

Tobii Technology AB. *Tobii T60 and T120 Eye Tracker User Manual*. Tobii Technology AB, 2008.

A. Tombros and M. Sanderson. Advantages of query biased summaries in information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 2–10, New York, NY, USA, 1998. ACM.

M. Trappett, S. Geva, A. Trotman, F. Scholer, and M. Sanderson. Overview of the INEX 2011 Snippet Retrieval Track. In *INEX 2011 Workshop*, pages 228–237, 2011.

Y. Tsegay, S. Puglisi, A. Turpin, and J. Zobel. Document compaction for efficient query biased snippet generation. In *Proceedings of the 31th European Conference on IR Research*, pages 509–520. Springer Berlin / Heidelberg, 2009.

A. Turpin, F. Scholer, and B. V. Billerbeck. Examining the pseudo-standard web search engine results page. In *Proceedings of the 11th Australasian Document Computing Symposium*, 2006.

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

A. Turpin, F. Scholer, K. Jarvelin, M. Wu, and J. S. Culpepper. Including summaries in system evaluation. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 508–515, New York, NY, USA, 2009. ACM.

L. Vanderwende, H. Suzuki, C. Brockett, and A. Nenkova. Beyond SumBasic: Task-focused summarization with sentence simplification and lexical expansion. *Information Processing & Management*, 43(6):1606–1618, 2007.

R. Varadarajan and V. Hristidis. A system for query-specific document summarization. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 622–631, New York, NY, USA, 2006. ACM.

I. Varlamis and S. Stamou. Semantically driven snippet selection for supporting focused web searches. *Data & Knowledge Engineering*, 68:261–277, 2009.

A. A. Verstak and A. Acharya. Generation of document snippets based on queries and search results. U.S. Patent 8.145.617, 2012.

E. M. Voorhees. Query expansion using lexical-semantic relations. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 61–69. Springer-Verlag New York, Inc., 1994.

E. M. Voorhees. The TREC-8 Question Answering Track Report. In *1999 Text REtrieval Conference*, pages 77–82, 1999.

C. Wang, F. Jing, L. Zhang, and H. Zhang. Learning query-biased web page summarization. In *Proceedings of the sixteenth ACM conference on Information and knowledge management*, pages 555–562, New York, NY, USA, 2007. ACM.

R. W. White, J. M. 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.

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

J. Xu and W. B. Croft. Query expansion using local and global document analysis. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 4–11, New York, NY, USA, 1996. ACM.

J. Xu and W. B. Croft. Improving the effectiveness of information retrieval with local context analysis. *Journal ACM Transactions on Information Systems*, 18(1):79–112, 2000.

G. Yule. *The study of language.* Cambridge University Press, 4th. edition, 2010.

H. P. Zhang, H. B. Xu, S. Bai, B. Wang, and X. Q. Cheng. Experiments in TREC 2004 Novelty Track at CAS-ICT. In *The 13th Text Retrieval Conference (TREC-2004)*, 2004.

J. Zobel and A. Moffat. Exploring the similarity space. *SIGIR Forum*, 32(1):18–34, 1998.