Technical Report

UCAM-CL-TR-744
ISSN 1476-2986

Number 744

UNIVERSITY OF CAMBRIDGE
Computer Laboratory

Citation context analysis for information retrieval

Anna Ritchie

March 2009

15 JJ Thomson Avenue
Cambridge CB3 0FD
United Kingdom
phone +44 1223 763500

http://www.cl.cam.ac.uk/

This technical report is based on a dissertation submitted June 2008 by the author for the degree of Doctor of Philosophy to the University of Cambridge, New Hall.

Technical reports published by the University of Cambridge Computer Laboratory are freely available via the Internet:

http://www.cl.cam.ac.uk/techreports/

ISSN 1476-2986

Abstract

This thesis investigates taking words from around citations to scientific papers in order to create an enhanced document representation for improved information retrieval. This method parallels how anchor text is commonly used in Web retrieval. In previous work, words from citing documents have been used as an alternative representation of the cited document but no previous experiment has combined them with a full-text document representation and measured effectiveness in a large scale evaluation.

The contributions of this thesis are twofold: firstly, we present a novel document representation, along with experiments to measure its effect on retrieval effectiveness, and, secondly, we document the construction of a new, realistic test collection of scientific research papers, with references (in the bibliography) and their associated citations (in the running text of the paper) automatically annotated. Our experiments show that the citation-enhanced document representation increases retrieval effectiveness across a range of standard retrieval models and evaluation measures.

In Chapter 2, we give the background to our work, discussing the various areas from which we draw together ideas: information retrieval, particularly link structure analysis and anchor text indexing, and bibliometrics, in particular citation analysis. We show that there is a close relatedness of ideas between these areas but that these ideas have not been fully explored experimentally. Chapter 3 discusses the test collection paradigm for evaluation of information retrieval systems and describes how and why we built our test collection. In Chapter 4, we introduce the ACL Anthology, the archive of computational linguistics papers that our test collection is centred around. The archive contains the most prominent publications since the beginning of the field in the early 1960s, consisting of one journal plus conferences and workshops, resulting in over 10,000 papers. Chapter 5 describes how the PDF papers are prepared for our experiments, including identification of references and citations in the papers, once converted to plain text, and extraction of citation information to an XML database. Chapter 6 presents our experiments: we show that adding citation terms to the full-text of the papers improves retrieval effectiveness by up to 7.4%, that weighting citation terms higher relative to paper terms increases the improvement and that varying the context from which citation terms are taken has a significant effect on retrieval effectiveness. Our main hypothesis that citation terms enhance a full-text representation of scientific papers is thus proven.

There are some limitations to these experiments. The relevance judgements in our test collection are incomplete but we have experimentally verified that the test collection is, nevertheless, a useful evaluation tool. Using the Lemur toolkit constrained the method that we used to weight citation terms; we would like to experiment with a more realistic implementation of term weighting. Our experiments with different citation contexts did not conclude an optimal citation context; we would like to extend the scope of our investigation. Now that our test collection exists, we can address these issues in our experiments and leave the door open for more extensive experimentation.

Contents

Chapter 1

Introduction

This thesis belongs in the field of information retrieval (IR), which may be loosely described as the finding of documents to satisfy a user's information need. We are particularly concerned with scientific literature search, i.e., retrieval of scientific research papers, rather than web pages or other types of documents. Literature search has always been an important part of science research: researchers disseminate their own ideas and results by publishing their work and also consult the literature to stay aware of what else is going on in their field.

Originally, literature search was a largely manual enterprise, involving libraries, reference librarians, card indexes and printed books and journals. Nowadays, automated literature search increasingly becomes the prevalent method of finding relevant research, as an ever growing abundance of literature is made available electronically, through digital research archives and academic publishers' as well as individual researchers' web sites. The advent of online publications, i.e., those which are only ever made available on the World Wide Web (henceforth, the Web), is a particular motivation for developing automated search. The vast, burgeoning quantity of material available means there is an increased danger that relevant work is never discovered, leading to duplication of work and effort. Improving automated literature search is therefore an important, ongoing area of research.

Traditionally, there are two main methods for searching the literature, called *subject indexing* and *citation indexing*. In subject indexing, descriptions of the subject matter of the individual documents are used to try to locate them. In citation indexing, the citations between documents are recorded and documents are located by following these links. Both of these approaches to search have undergone automation. Nowadays, citation indexes are generated automatically from machine-readable documents and are themselves published electronically, often as part of online search tools, rather than being printed on paper. Likewise, subject indexing is largely done automatically, rather than by human indexers choosing keywords.

Subject indexing and citation indexing are orthogonal, complementary approaches to searching the literature. The former deals with document-internal properties; the latter with inter-document relationships. Searching directly by citation, as in a citation index, requires a seed document from which to begin the search, rather than just an idea or statement of what the search is about. However, the possibilities for using citations for literature search extend beyond straightforward citation hopping. A citation encodes far more than a simple pointer from one document to another. The citation process is a complex phenomenon that is interesting to researchers in various disciplines, for disparate reasons. The frequencies and patterns of citation, the functions of citation and the motivations for citing have all been studied.

Furthermore, research in Web IR has produced a great wealth of link methods that could naturally be applied to the citation link structure of literature for searching. These link methods, moreover, have been combined effectively with methods based on the internal content of web pages.

The time is ripe for integrating citation- and subject-based literature search. There has been a recent resurgence of interest in citations as a topic of research, both in IR and further afield. Automated methods for processing and analysing citations are being developed, building on the foundations of earlier analytical citation research. These methods combined with automated indexing and an abundance of readily available machine-readable documents make it practicable to more freely experiment with integrating citation information with existing search technology.

The work presented in this thesis draws together ideas from IR and citation analysis. In Chapter 2, we expand on the motivation for our work, give an overview of related research and try to position our work therein. Chapter 3 gives a theoretical account of test collections, the experimental paradigm used to evaluate IR systems, before describing how and why a test collection was created for the evaluation presented in this thesis. Chapters 4 and 5 describe, respectively, the document collection around which our test collection was built and how those documents were processed in order to conduct our experiments. In Chapter 6, we summarise our experimental set-up and results. Finally, in Chapter 7, we conclude by summarising the contributions of this thesis and discussing potential directions for future work.

Chapter 2

Background and motivation

The term information retrieval (IR) was coined by Mooers (1951) and made popular by Fairthorne (1956), and denotes the finding of information to satisfy a user's information need. As a field of study, IR lies at the boundary between information science and computer science. Citation analysis has been an important line of research in information science since the 1960s and citation information has, accordingly, been used in IR. Indeed, most early IR systems were designed around bibliographic databases. In this chapter, we map the motivation for this thesis work, discussing how citations are traditionally used in IR, before exploring broader work in both citation analysis and then IR. We finally describe our own experiments and how they draw together related ideas from these areas. First, we give an overview of the fundamental concepts in IR.

2.1 IR basics

The task of IR systems is to find information that satisfies a user's information need. Nowadays, IR is typically synonymous with *document retrieval*. Thus, the task becomes to find documents with information content that satisfies the user's need. Written, natural language documents are of special significance in IR, since natural language is the normal manner of communication for people. Accordingly, document or information retrieval can often mean *text retrieval*. Nevertheless, a document can be any item that conveys information, such as image, video or audio files. It is typical for non-textual documents to be augmented with textual representations (e.g., captions for images) and for text retrieval methods to be used to search for the non-textual documents. Even in proper text retrieval, the full-text documents themselves need not necessarily be the objects that are stored and searched by the system: they may be represented by *document surrogates*, such as their titles and abstracts. For the purposes of this discussion, though, we will use the term *document* to mean the object inside the retrieval system.

Information is an indefinite, intangible quantity and, thus, both document information content and user information needs are unobservables. Therefore, the usual simplifying presumption in IR is that users want to find out about a *topic* and so the retrieval system must find documents that are *about* the topic or, in other words, that are *relevant* to the user's need. Users must express their information need somehow, whether as a natural language *request* or a *query*, defined within system constraints. The retrieval system's aim is then to capture the relevance relation by establishing a matching relation between two expressions of information: the query and a document. The central issue in IR is ensuring that the documents that the system matches with the query are about the same topic. The retrieval process is divided into two main opera-

tions: *indexing* and *searching*.

Indexing is the process of creating *representations* of documents and queries upon which to conduct the matching. Indexing languages should be designed so that, if the respective representations of a query and document match, the document is relevant to the query and, conversely, if the document is not relevant to the query, their representations should not match. Therefore, an indexing language should overcome surface differences between documents and queries that are about the same topic, while simultaneously being robust to surface similarities between documents and queries that are *not* about the same topic.

The individual items in the indexing language are called index *terms*. The indexing language may be a finite set of predetermined terms, such as a library classification or a thesaurus of subject-specific terms, or, typically nowadays, may be natural language, i.e., the words from the documents are used as the indexing language. In this case, the document text will typically be manipulated to arrive at the final index terms. For instance, common words that do not contribute to the meaning of a document may be removed (*stopping*) and words may be reduced to their linguistic stems (*stemming*) to conflate semantically related words.

The second operation in retrieval is searching. The core subtasks in searching are *matching*, where the system establishes what a document representation and a query representation have in common, and *scoring*, where the system assigns a score to the match that reflects the strength or 'goodness' of the match. These subtasks are not necessarily independent of each other, depending on the *retrieval model* that the system uses, where a retrieval model is a formal specification of the entire retrieval process. There are very many retrieval models. We consider a few main classes of model here: the *Boolean model*, the *vector space model*, the *probabilistic model* and *language modelling based models*.

A simple retrieval model is the Boolean model, where the query terms are connected by Boolean logic operators (AND, OR and NOT), the presence of a query term in a document is sufficient for a match on that term and only documents whose terms satisfy the entire Boolean query expression are matched and retrieved. In this case, scoring is a binary decision: documents are either relevant or not, and are not scored against each other.

The majority of other retrieval models rank retrieved documents according to how well they score for the query relative to each other. Usually in such models, it is not necessary for all query terms to be present in a document in order for it to be retrieved and some sort of term weighting scheme is applied when calculating scores to reflect the relative importance or distinctiveness of different terms. For instance, the classic TF*IDF weight is designed to prioritise terms that occur often in a document (i.e., have a high *term frequency*) and that are rare in the overall document collection (i.e., have a high *inverse document frequency*). Then, the weight of a term t in document d from document collection D is thus calculated:

$$
\begin{aligned}
TF * IDF_{t,d,D} &= TF_{t,d}.IDF_{t,D} \\
TF_{t,d} &= freq_{t,d} \\
IDF_{t,D} &= \log \frac{|D|}{n_{t,D}} \\
n_{t,D} &: \text{number of documents in } D \text{ with term } t
\end{aligned}
$$

In the vector space model (Salton, Wong & Yang 1975), queries and documents are represented as vectors in a high-dimensional space, where each term in the indexing language is a

dimension. The coordinate values in a vector may be binary (to indicate presence/absence of that term) or may be weighted, e.g., by TF*IDF. Relevance is then modelled by document-query similarity: documents closer to a query in the vector space are taken to be more relevant and scored higher. The distance between two vectors is defined by some proximity measure, such as the cosine similarity measure, i.e., the normalised dot product of the two vectors.

Probabilistic retrieval models are based on the probability ranking principle, which says that if a retrieval system ranks documents by decreasing probability of relevance to the user, it will be the most effective it can be to its users (Cooper 1972, cited after Robertson 1977). Therefore, these models estimate the probability of a document being relevant for a given query and then rank documents by their probabilities.

A more recent class of retrieval model uses language modelling techniques to incorporate more sophisticated statistics about the language in queries and documents into the scoring. For instance, given a query, documents may be ranked by the probability that the language model calculated from that document would generate the sequence of terms in the query. Similarly, a language model may be calculated over the query and each of the documents; then, the similarity between the query and document models are used for scoring.

2.2 Citations in IR

Citations signify intellectual linkages between academic works and this *link structure* can be followed, backwards as well as forwards, to search for relevant papers; this is the basic premise of citation indexing. Citation indexes were early established as an important tool for searching in collections of academic literature (Garfield 1979). Thus, citations provide information scientists with an alternative type of information to that inside the individual works themselves; citation indexing is an orthogonal, complementary alternative to subject indexing, for connecting users with relevant academic works. Nowadays, academic literature is increasingly available on the Web and automated citation indexing has been combined with online search in tools such as Citeseer[1](Lawrence, Bollacker & Giles 1999) and Google Scholar[2].

In addition to their use as a direct search tool, citation indexes provide statistical data about citations in the indexed body of work, from which various citation analyses can be conducted. For instance, two core citation analyses are bibliographic coupling (Kessler 1963), where documents are said to be coupled if they share one or more references, and co-citation analysis (Small 1973), where the similarity between documents A and B is measured by the number of documents that cite both A and B. The theory behind bibliographic coupling is that documents that are similar in subject have similar references; the theory behind co-citation analysis is that documents that are similar are more likely to be cited by the same other documents. These principles each provide a means of quantifying document similarity or relatedness using citations. Consequently, both bibliographic coupling and co-citation analysis have commonly been put to use in IR over the years. There is, in fact, a tradition in IR of using methods based on statistical citation information, which continues today. For instance, Strohman, Croft & Jensen (2007) use co-citation data as one feature in a system that, given a document as a 'query', retrieves documents to be recommended for citation by that document; Meij & de Rijke (2007) use citation counts to estimate prior probability of document relevance in a language model retrieval framework; Fujii (2007) experiments with a variant of PageRank calculated from citation counts for patent retrieval. (See Section 2.4 for PageRank.) This statistical focus in IR is

[1] `http://citeseer.ist.psu.edu/`

[2] `http://scholar.google.com/`

not, however, a characteristic of citation analysis in general: citation analysis is a broad and extensive topic of research, with many qualitative studies.

2.3 Citation analysis

Citation analysis is the study of the relationship between (part of) a citing document and (part of) a cited document that the citation implies. Several disciplines conduct research in citation analysis, such as applied linguistics and history and sociology of science, as well as information science. In information science, the subfield to which citation analysis belongs is *bibliometrics*: the quantititative study of writings in the aggregate. The term 'quantitative' here does seem to imply that information scientists are, indeed, concerned only with the numbers of citations that occur but, despite its definition, bibliometrics has born many qualitative citation studies, as we shall see. That said, one of the most notable contributions of information science to citation analysis is certainly quantitative: the development of bibliometric measures and tools for science management, i.e., statistical measures used for qualitative evaluation of individuals, institutions, publications and even countries in science, and tools that calculate and present these measures. For instance, *journal impact factor* is a measure of the frequency with which the journal's average article is cited; citation counts and rates of individuals/institutions are used to gauge their scientific productivity. The Science Citation Index[3] and Journal Citation Reports[4] are notable examples of science management tools. Garfield (1979) notes that, although citation indexes were developed primarily for bibliographic purposes, science management may in fact be their most important application. However, there is a long history of controversy over the use of citation counts as a measure of quality and many theoretical objections have been raised, which Garfield goes on to discuss. For instance, citation counts may be inflated by self-citation, how to relate the citation rate of co-authored papers to individual authors is not clear and articles in prestigious journals may be cited more than articles of equal quality due to the journal's visibility. Probably the most prominent criticism is that citation analysis based on raw citation counts ignores the underlying reasons for the citation; works may be cited in refutation or as negative examples. Thus, not all citations are positive endorsements of the cited work. Such criticisms have prompted many citation studies, in order to gain a better understanding of the citation process and the validity of citation analysis for quality assessment; Liu (1993) presents a thorough, relatively recent review.

Research in citation analysis may be categorised in a number of ways. Liu (1993), for example, labels studies by their research objectives, giving five labels (which are not mutually exclusive): to enhance citation indexes, to describe the manifest functions of citations, to assess the quality of citations, to define concepts attributed to the citing work by the citing work and, lastly, to examine the underlying motives for citing. Alternatively, citation studies can be categorised by methodology. Taking this approach, work may first be divided into two broad categories: roughly speaking, whether the text of the citing document is the object of study or not. The majority of work falls into the first category, called *citation context analysis* (Small 1982), where 'context' means the textual passage or statement that contains the citation. The second category seems to exclusively contain work on *citer motivations*, i.e., examining the motives that authors have for citing, which are outside the text. Brooks (1985) is cited as the first study of real authors and their motives for citing. Through surveys and interviews, this work identified seven motivational variables, including *persuasiveness*, *reader alert* and both *positive* and

[3] http://scientific.thomson.com/products/sci/

[4] http://scientific.thomson.com/products/jcr/

negative credit, and found persuasiveness to be the main motivating factor, i.e., the citing author is marshalling earlier work in order to persuade readers of the quality of his or her own claims.

The scope of citation context analysis, as loosely defined above, is very broad. Small (1982) subdivides citation context analyses into those which use the citing text to abstractly classify citations and those which use it to identify the particular, concrete topics that are being attributed to the cited work, though the two approaches are not always entirely distinct. The first may be unambiguously called *citation classification* and is principally concerned with the relationship between the citing and cited document. The second is concerned with the topical content of the cited document and has been called *citation content analysis* and *content analysis of citation contexts*. This is confusable with the term *content citation analysis* (used by, e.g., Swales 1986, Teufel 1999), which comes from *content analysis*, a standard methodology in the social sciences for studying the content of communication (Krippendorff 2004), but means something more general[5]. Citation classification schemes define a taxonomy with which to encode the relationship between the citing and cited documents. By manifesting these relationships, citation classification allows the patterns of citation to be studied in finer detail and, furthermore, makes it possible for citation analytic techniques to discriminate between types of citations. Moravcsik & Murugesan (1975) present one of the earliest classification schemes, intended to improve understanding of citations and, specifically, the extent to which citation counts can be used for various purposes in science policy. In this study, 74 citations to the same book from 30 physics articles were manually classified according to four dichotomous categories: *conceptual/operational*, *organic/perfunctory*, *evolutionary/juxtapositional* and *confirmational/negative*. This scheme was modified in later work in order to make the categories more appropriate to a non-scientific field and easier to code (Swales 1986). The problem of difficulty and subjectivity in coding is common to most classification schemes; most schemes were only ever intended to be executed by the scheme's creators, on a small number of sample papers, in order to make generalisations about citation practices from intensively studying a few examples. This precludes their being applied automatically, limiting their practical use in an age when large bodies of literature are available in electronic form. More recently, schemes have been developed with automatic classification in mind. Nanba & Okumura (1999), for example, devised a classification with only three categories (*compare*, *based-on* and *other*) and manually created rules based on cue words/phrases to be applied automatically by a system to support writing domain surveys. Teufel, Siddharthan & Tidhar (2006) adapted an earlier manual classification scheme (Spiegel-Rösing 1977) in order to make the annotation task more objective, systematic and, therefore, reliable among humans, so that an automatic system can better replicate the procedure. A machine learning classifier is trained on human annotated papers, using features like cue phrases and location within the paper, section and paragraph. Teufel et al. observe that citation classification has strong ties with rhetorical discourse structure analysis and note its consequent usefulness in tasks like text summarisation.

In citation content analyses, the explicit, contentful words that the citing author uses to describe the cited work are the object of study. Citation markers are usually introduced purposefully alongside some descriptive reference to the cited document and these descriptions

[5]Holsti (1969) defines content analysis as 'any technique for making inferences by objectively and systematically identifying specified characteristics of messages'. Thus, *content citation analyses* are content analyses of citing documents and, strictly speaking, *all* citation context analyses are content analyses of citation contexts. The 'content' in *citation content analysis* refers to the topical content of the *cited* document, rather than to the status of the citing document as the content under analysis.

hyperlink: The Google search engine...

citation: 'Dictionaries can be constructed in various ways - see *Watson (1993a, 1995)* for a taxonomy of (general) finite-state automata construction algorithms.'

Figure 2.1: Similarity between hyperlinks and citations.

may be taken as summaries of what that document is about or, at least, what the citing author thinks is important about the work. In other words, citation contexts are rich in keywords for the cited work. O'Connor (1982), therefore, argued that noun phrases from these *citing statements* are useful as subject index terms and should be used to augment an existing document representation to improve retrieval effectiveness. Using this same idea of keyword-richness in citation contexts, more recently, Schneider (2004) investigated using citation contexts in semi-automatic thesaurus construction. Also, Nakov, Schwartz & Hearst (2004) automatically extract paraphrases of facts about a cited paper from multiple citations to it, with the eventual aim of using these to automatically create summaries of the cited paper. This is allied with Small's (1978) notion of cited works as *concept symbols*, whereby a work may come to be repeatedly and consistently cited to represent a specific idea or topic, using descriptions that converge on an almost fixed terminology for that topic, such that the cited work eventually becomes synonymous with the topic.

However, while the potential usefulness of citation context analysis has been noted in relation to many modern applications – we have already come across text summarisation, thesaurus construction, scientific authoring tools and rhetorical discourse structure annotation – work in IR has continued to focus on statistical citation data, like citation counts. This segregation of ideas and methods has been observed on a much greater scale: White (2004) discusses three separate fields with a tradition of citation analysis (history and sociology of science, applied linguistics and information science) and notes great interdisciplinary differences in how citations are used. Moreover, White gives pointed examples of how methods from one discipline could be put to good use in another, and urges more dissemination and sharing of knowledge between disciplines. In this thesis work, we draw together ideas from IR and the sort of citation context analysis already applied in other fields.

We have seen that citations are the formal, explicit linkages between papers that have particular parts in common and that this *link structure* is encoded in citation indexes so that users can follow links to try to find relevant work. We now observe that the link structure formed by citations is analogous to that of the Web, where the links are hyperlinks between web pages. In Pitkow & Pirolli's (1997) words, 'hyperlinks ... provide semantic linkages between objects, much in the same manner that citations link documents to other related documents.' In this analogy, the textual marker that denotes a citation is the value of the `href` attribute of the HTML hyperlink, while the citation context, i.e., the descriptive reference to the cited work, is the anchor text, i.e., the text enclosed in the ⟨a⟩ (anchor) tags of the HTML document (see Figure 2.1). The link structure of the Web, including anchor text, has been studied extensively in IR and exploited to advantage in some retrieval tasks. In the following section, we review this work.

2.4 Link structure and anchor text in IR

Web pages are often poorly self-descriptive and this means that indexing only the content of the pages (akin to traditional subject indexing for literature search) will often give poor results. Kleinberg (1999) notes that `www.harvard.edu` is not the page that contains the term 'Harvard' most often, that search engine home pages like Yahoo! and Google do not contain the term 'search engine' and, likewise, that there is no reason to expect car manufacturers to label themselves with 'car manufacturer' on their home pages. Researchers in IR have turned to the link structure of the Web as an alternative source of information about which pages should be retrieved for a given query.

Kleinberg (1999) presents the Hypertext Induced Topic Selection (HITS) ranking algorithm[6], which relies solely on hyperlinks to discover *authoritative* information sources from the output of a search engine on a given topic. The motivation behind this algorithm is that hyperlinks encode considerable latent human judgement: by linking to a page, the author of a hyperlink confers some *authority* on that page. Each page, therefore, has an authority value, that estimates the value of the content of the page, and a hub value, that estimates the value of its links to other pages. These values are defined mutually recursively: the authority value of a page is the sum of the normalised hub values that point to that page and the hub value is the sum of the normalised authority values of the pages it points to. The values are computed iteratively over a relatively small subgraph of the Web which, by design, should contain many relevant pages and many of the strongest authorities on the query topic, e.g., the output of a search engine in response to the topic. The pages are then re-ranked by their authority values.

Brin & Page (1998) describe PageRank, a measure of web page *importance*, based on weighted, normalised forward- and backward-link counts. The PageRank of a page is defined recursively, depending on the number and PageRanks of all pages that link to it. The intuitive justification here is that a page can have a high PageRank (i.e., be important) if many pages point to it or if an important page points to it. Retrieved pages can then be re-ranked by their PageRank. In contrast to HITS authority values, PageRanks are calculated from the entire graph and, thus, can be calculated offline, increasing query-time efficiency.

Hotho, Jäschke, Schmitz & Stumme (2006) adapted PageRank to the task of retrieving resources within social resource sharing systems, such as videos in YouTube[7] and photographs in Flickr[8]. In these systems, the users create lightweight conceptual structures called *folksonomies* (from 'folk' and 'taxonomy') by assigning free-text tags to resources in order to classify them however they wish. This structure can be converted into an undirected graph and an adaptation of PageRank applied to it, such that resources tagged with important tags by important users become important. Fujii (2007) applies the PageRank idea to patent retrieval, using counts of how often patents are cited by other patents to re-rank them.

Link structure techniques like HITS and PageRank are applications of *social network analysis* to the Web. Social network analysis is widely used in the social and behavioural sciences, as well as in economics, marketing and industrial engineering, and tries to express social (or

[6] There is a poetic irony here. Kleinberg (1999) notes the problem of poorly self-descriptive web pages, yet suffers from an analogous problem: the paper is the standard reference for HITS but the name HITS was only attributed to the algorithm after the paper was written. So, while the paper has taken on the status of a concept symbol for 'HITS', to use Small's (1978) terminology, and is commonly cited in conjunction with this term, it does not contain the term at all.

[7] `http://www.youtube.com`

[8] `http://www.flickr.com`

political or economic) environments as patterns of regularities in the relationships among the interacting units (Wasserman & Faust 1995). This has found other technological applications besides IR, such as reputation management in electronic communities, where the community is modelled as a social network and aims to avoid interaction with undesirable members (like spammers) by combining hard security techniques with social mechanisms, i.e., trust (Yu & Singh 2000). Agrawal, Rajagopalan, Srikant & Xu (2003) use link structure methods to mine newsgroup discussions, in order to classify hypertext documents as for or against an issue.

Social network algorithms applied to the Web are principally concerned with the presence of hyperlinks; with the quantities and patterns they occur in. These ideas have also been extended to look further than the raw hyperlink, at the text associated with the hyperlink inside the linking web page. While web pages may not describe themselves very well, pages which contain hyperlinks often describe the pages they point to. More specifically, the anchor text of hyperlinks is often a description of the pointed-to page. Thus, beginning with the WWW Worm search engine (McBryan 1994), there is a trend of propagating anchor text along its hyperlink to associate it with the linked page; Figure 2.2 illustrates this method. Google, for example, indexes anchor text for the linked page (Brin & Page 1998) and this is believed to make a significant contribution to its retrieval effectiveness. In competitive evaluations, using anchor text has been shown to be very effective for navigational search tasks (e.g., site finding), though not topic relevance tasks (i.e., retrieving pages relevant to a query topic) (Craswell, Hawking & Robertson 2001). Anchor text is not unique in being descriptive of the linked page, however. Bharat & Mihaila (2001) generalise the notion to *qualifying text* and include, e.g., titles and headers in the linking page, as well as anchor text, in a ranking algorithm that only uses links deemed relevant to the query, i.e., those whose qualifying text contains query terms. Chakrabarti et al. (1998) look for topic terms in a fixed window of text around the `href` attribute of hyperlinks and weight that link accordingly, in an augmented version of HITS. Conversely to the anchor text idea, Marchiori (1997) recursively augments the textual content of a page with all the text of the pages it points *to*; the idea here is that, by providing access to it, the linking page in some sense 'contains' the information in the linked page. Marchiori implemented this idea as a re-ranking step on the output of popular search engines and conducted a user-centric evaluation, showing that users generally preferred the re-rankings to the original rankings.

2.5 Indexing citation contexts

We have seen how the descriptiveness of anchor text (and associated text) has been put to use in Web IR, by indexing it along with the page it describes. We have also seen how hyperlinks and anchor text may be likened to citations and their associated descriptive text; their contexts. The aim of the work presented in this thesis is to explore whether indexing the contexts of citations to a paper in combination with the paper itself can improve the retrieval performance achieved when only the paper is indexed. Some work has been done in this area but no previous experiments have used both citing and cited papers to the fullest possible extent.

The idea of taking the terms that citing authors use to describe a cited work and using them as index terms for the cited paper predates the analogous use of anchor text in Web IR. O'Connor (1982) motivated the use of words from *citing statements* as additional terms to augment an existing document representation. Though O'Connor did not have machine-readable documents, procedures for 'automatic' recognition of citing statements were developed and manually simulated on a collection of chemistry journal articles. Proceeding from the sentence in which a citation is found, a set of hand-crafted, mostly sentence-based rules were applied to

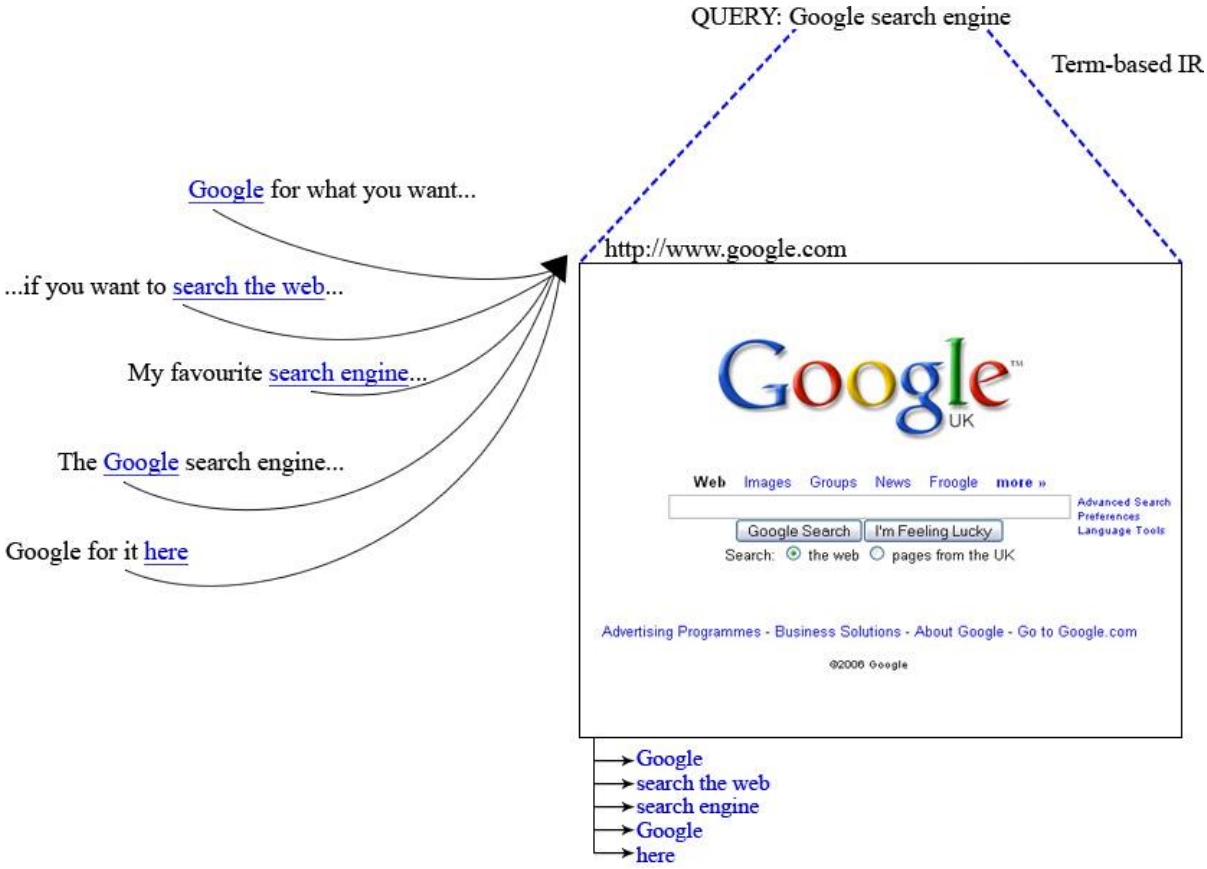


Figure 2.2: Propagating anchor text for web page indexing.

select the parts of the citing paper that conveyed information about the cited paper. The selected statements (minus stop words) were added to an existing representation of the cited documents, comprising human index terms and abstract terms, and a small-scale retrieval experiment was performed. A 20% increase in recall was found using the citing statements in addition to the existing index terms, though in a follow-up study on biomedical papers, the increase was only 4%; O'Connor attributes this drop to a lower average number of citing papers in the biomedical domain (O'Connor 1983). O'Connor concluded that citing statements can aid retrieval but notes the inherent difficulty in identifying them. Some of the selection rules were only semi-automatic (e.g., required human identification of an article as a review) and most relied on knowledge of sentence boundaries. Sentence boundary detection is a non-trivial computational problem in itself, particularly in scientific text, which is rife with formulae, unit abbreviations and textual place holders. Nowadays, tools for sentence boundary detection are widely available and a solution to the problem of automatically identifying review articles is also within reach: Nanba & Okumura (2005) present a method for detecting survey articles in a multilingual database.

More recently than O'Connor's studies, Bradshaw (2003) implemented *Reference-Directed Indexing* (RDI), whereby a scientific document is indexed by the text that refers to it in citing documents, instead of by the text in the document itself, as is typical in IR. The theory behind RDI is that, when citing, authors describe a document in terms similar to a searcher's query for the information it contains. Thus, Bradshaw hypothesises that this *referential text* should contain good index terms for the document and shows an increase in precision over retrieval by the document terms alone, using a standard vector space model implementation; 1.66 more

relevant documents are retrieved in the top 10 in a small evaluation on 32 queries.

However, a number of issues may be raised with RDI. Firstly, referential text is extracted using CiteSeer's *citation context*; a window of around one hundred words around the citation. This method is simplistic: the words that are definitely associated with a citation are variable in number and in distance from the citation, so a fixed window will not accurately capture the citation terms for all citations. Indeed, Bradshaw states the difficulty in extracting good index terms automatically from a citation, echoing O'Connor. Bradshaw's experiment is limited by the use of the Citeseer data and he does not compare with any alternatives to the fixed window. Secondly, RDI only indexes referential text and not the text from the documents themselves, so a document must be cited at least once (by a document available to the indexer) in order to be indexed at all. This has particular consequences for recently published documents, as it takes time for works to be disseminated, responded to and eventually cited. RDI makes no fall-back provision for uncited documents; Bradshaw's evaluation excluded any documents that were not cited and does not disclose how many of these there were.

Dunlop & van Rijsbergen (1993) investigated a similar technique with a different application in mind (i.e., retrieval of non-textual documents, such as image, sound and video files). Dunlop's retrieval model uses clustering techniques to create a description of a non-textual document from terms in textual documents with links to that document. In order to establish how well these descriptions represent the documents, the method was applied to textual documents, indeed, to the CACM test collection, where the documents are abstracts from scientific papers and the links between documents are citations. The experiment compared retrieval performance using the cluster-based descriptions against using the documents themselves; the cluster-based descriptions achieved roughly 70% of the performance achieved using the document content. Again, Dunlop did not measure performance using the cluster-based descriptions in combination with the document content. Additionally, the text taken from the citing papers was simply the whole abstract and not specifically the text used in association with the citations.

2.6 Thesis goal and feasibility study

Thus, retrieval effectiveness using the combination of terms from citing and cited documents has not previously been fully measured, nor compared with the use of terms from only cited or citing papers. This thesis will therefore present the results of retrieval experiments novelly using the combination of document and citation terms, compared to using the document terms alone, as illustrated in Figure 2.3. Our goal is to test the hypothesis that an existing document representation comprised of the full text of the document will be enhanced by adding to it terms from citations to the document.

As a preliminary test-of-concept experiment for this research, we studied 24 citations to one paper entitled *The Mathematics of Statistical Machine Translation: Parameter Estimation*, from the Computational Linguistics journal[9]. Ritchie, Teufel & Robertson (2006*b*) present this study in full. We manually identified the words from around those citations that specifically referred to the paper. There is no explicit anchor text in scientific papers, unlike in web pages, where there are HTML tags to delimit the text associated with a link. Identifying which words are associated with a citation is an interesting, complex problem, which has been discussed in depth (O'Connor 1982, Ritchie et al. 2006*b*). For example:

- The amount of text that 'belongs' to a citation can vary greatly, so a fixed window will

[9] `http://acl.ldc.upenn.edu/J/J93/J93-2003.pdf`

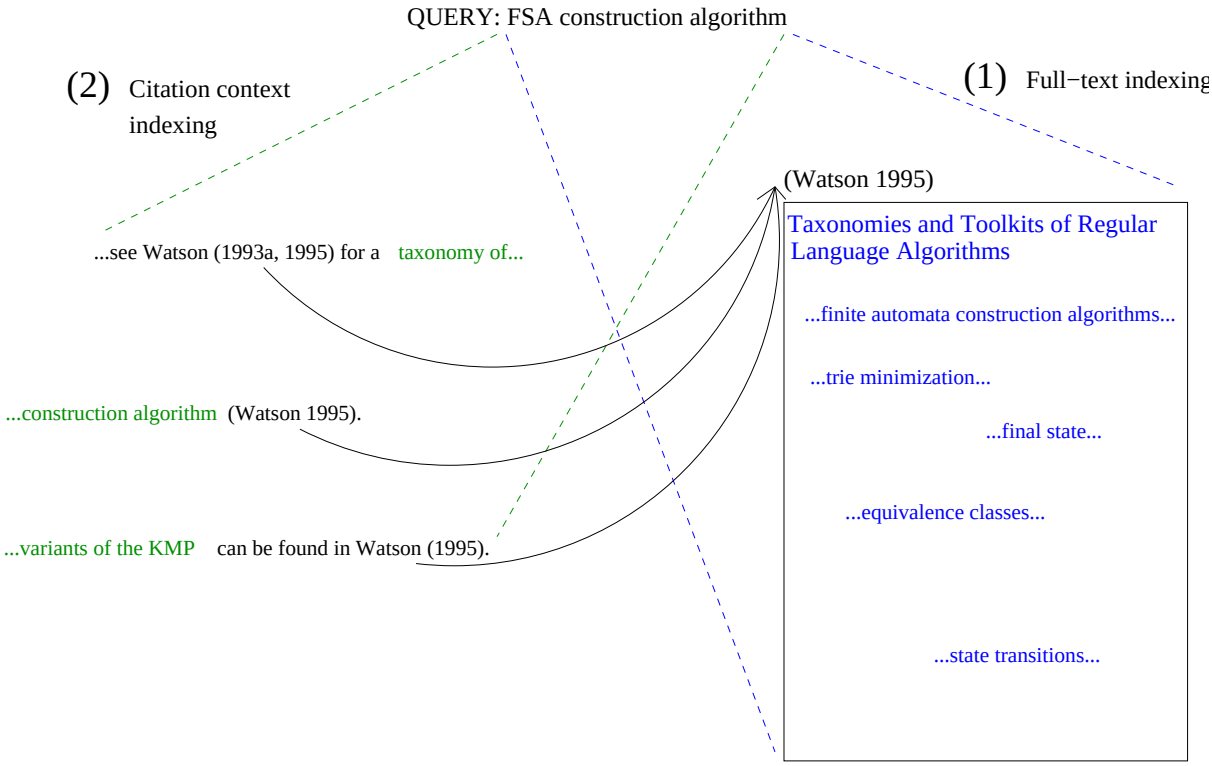


Figure 2.3: Combining (1) citation context indexing and (2) full-text indexing.

. . . model to the chunk.
Here, the back-off model is applied only to the part that failed to get translation candidates.
3.1 Learning Chunk-based Translation
We learn chunk alignments from a corpus that has been word-aligned by a training toolkit for word-based translation models: the Giza++ (Och and Ney, 2000) toolkit for the **IBM models (Brown et al., 1993)**. *For aligning chunk pairs, we consider word(bunsetsu/eojeol) sequences to be chunks if they are in an immediate dependency relationship in a dependency tree. To identify chunks, we use a word-aligned corpus, in which source language sentences are annotated with dependency parse trees by a dependency parser (Kudo et al., 2002) and . . .*

Figure 2.4: Example citation from P05-1068 to J93-2003, with 'ideal' citation terms in **bold**.

not accurately capture the citation terms for all citations.

- Multiple citations in close proximity can interact with each other and affect the 'ownership' of surrounding words.
- Likewise, sentence boundaries (as well as paragraph and section boundaries) can indicate a change of ownership, as they signal topic shifts.

Figure 2.4 shows an example citation context from this study, giving a one hundred word window around the citation, with the words that we identified as specifically referring to the paper in bold font; the 'ideal' citation terms (in this case, *IBM models*). We then observed the effect of adding the ideal citation terms from all 24 citations to the document terms and noted four main points of interest.

Rank		TF*IDF	Term
Ideal	Doc		
1	1	351.73	french
2	2	246.52	alignments
3	3	238.39	fertility
4	4	212.20	alignment
5	5	203.28	cept
6	8	158.45	probabilities
7	9	150.74	translation
8	12	106.11	model
9	17	79.47	probability
10	18	78.37	models
11	19	78.02	english
12	21	76.23	parameters
13	24	71.77	connected
14	28	62.48	words
15	32	57.57	em
13	35	54.88	iterations
14	45	45.00	statistical
15	54	38.25	training
16	69	32.93	word
17	74	31.31	pairs
18	81	29.29	machine
19	83	28.53	empty
20	130	19.72	series

Table 2.1: Ideal citation term ranking by TF*IDF.

Firstly, there was overlap between the citation terms and important document terms. Table 2.1 gives the top 20 ideal citation terms ranked by their TF*IDF values in the original document, also giving the absolute rankings of these terms in the original document in the second column to give an indication of their importance relative to other terms in the document. From this we see that the five document terms with the highest TF*IDFs were also citation terms: *french*, *alignments*, *fertility*, *alignment* and *cept*. Thus, indexing the citation terms would reinforce the visibility of these important document terms.

Secondly, the TF*IDF values of intuitively important descriptor terms for the paper were substantially increased by adding the citation terms. Table 2.2 shows the ideal citation terms which moved most in the TF*IDF rankings as a result of adding the citation terms to the document terms. For instance, *ibm* is a distinctive term (with a high IDF) but appears only six times in the document (and not even from the main text but from authors' institutions and one bibliography entry) yet one of the paper's major contributions is the machine translation models it introduced, now standardly referred to as 'the IBM models'. Consequently, *ibm* occurred 11 times in the 24 citation contexts we studied and *ibm*'s TF*IDF is more than doubled when citation contexts are taken into account. This exemplifies how citation terms can sometimes better describe a document, in terms of what searchers might plausibly look for, as could be the case of a researcher looking for Kleinberg's HITS paper.

Thirdly, 20 of the citation terms did not occur at all in the document itself. Table 2.3 lists these 'new' non-zero TF*IDF terms and shows that many of them have high IDF values, indic-

Term	TF*IDF		Ideal
	Δ	Doc+ideal	rank Δ
ibm	24.24	37.46	28 → 20
generative	4.44	11.10	38 → 33
source	5.35	6.42	65 → 44
decoders	6.41	6.41	__ → 45
corruption	6.02	6.02	__ → 46
expectation	2.97	5.94	51 → 47
relationship	2.96	5.92	52 → 48
story	2.94	5.88	53 → 49
noisy-channel	5.75	5.75	__ →52
extract	1.51	7.54	41 → 38

Table 2.2: Term ranking changes (Ideal).

ating their distinctiveness, e.g., *decoders*, *corruption* and *noisy-channel*. Without the citation index terms, however, the paper would probably not be retrieved for queries with these terms.

Finally, we noted that some highly distinctive terms that did *not* refer to the paper would be wrongly picked up if a fixed window were used to extract the citation terms. For instance, Giza is the name of a toolkit that is used to train the IBM models and was developed after the models; Giza is, naturally, not mentioned in the IBM paper and should not be used to index the paper. However, since researchers who have used the IBM models often use Giza to train them, the term *giza* often appears in close proximity to citations to our example paper, as in Figure 2.4. Appendix A includes a list of the almost 400 'noisy' terms that were picked up by the fixed window but were not ideal citation terms. The appendix also includes the equivalent tables to Tables 2.2 and 2.3 looking at the fixed window terms, rather than ideal citation terms.

This case study highlights several interesting effects of using terms from around citations as additional index terms for the cited paper. However, it cannot answer questions about how successful a practical method based on these observations would be. The remainder of this thesis describes the setting up and execution of a full-scale experiment to shed more light on this question. A significant part of this research effort was spent creating a new test collection of scientific papers in order to evaluate our experimental method. In the following chapter, we describe why and how we built this collection, first giving a theoretical account of the test collection paradigm used in IR evaluation.

Term	TF*IDF
decoders	6.41
corruption	6.02
noisy-channel	5.75
attainable	5.45
target	5.24
source-language	4.99
phrase-based	4.92
target-language	4.82
application-specific	4.40
train	4.10
intermediate	4.01
channel	3.47
approaches	3.01
combinations	1.70
style	2.12
add	1.32
major	1.16
due	0.83
considered	0.81
developed	0.78

Table 2.3: New non-zero TF*IDF terms (Ideal).

Chapter 3

Test collection

This chapter describes the creation of a test collection of scientific papers, which was necessary for the evaluation of our retrieval experiments. Building a test collection is a difficult, time-consuming challenge and there are many issues associated with the process. In this chapter, we first give a theoretical overview of the test collection experimental paradigm, with a view to highlighting the issues that are relevant to our own test collection. We discuss how test collections are built, examining a number of specific examples to demonstrate the different approaches that may be taken. We then consider some existing test collections and show that they are unsuitable for our retrieval experiments. In Chapter 4, we will introduce the ACL Anthology and discuss the reasons for choosing it as the document collection for our experiments; now, we detail the work that went into creating a new test collection around this document collection. We conclude by describing the resultant test collection.

3.1 The test collection paradigm

The ultimate aim of the IR experimenter is to improve user satisfaction in the real world. How best to judge the performance of retrieval systems at this task is by no means straightforward, however, and evaluation has consistently been a source of contention over the years. It began in earnest with a National Science Foundation sponsored program in systems evaluation (Brownson 1960).

Tague-Sutcliffe (1992) enumerates the design decisions that the experimenter must make, including the choice between conducting a *laboratory* or *operational* test. Operational tests evaluate the performance of a retrieval system in a particular, real working environment; one or more complete systems are evaluated or compared, including their users, databases and search constraints. The experimenter hopes to improve the larger world by creating improvements on a smaller portion of the real world. They are generally agreed to be more realistic and are, thus, preferred. However, it is usually difficult to conduct tests in real operational environments, as such tests are likely to be disruptive to system operations. Also, achieving the level of control required for reliable results is usually not possible in an operational environment: there are very many parameters in real searches. There are different user types (e.g., Web surfers, patent lawyers, academics) of differing levels of expertise, age, gender and nationality. The nature of the searches can vary widely: queries can be simple or complex, broad or specific; searches may have a single, known relevant document or many, unknown relevant documents; the search task may require everything relevant to be retrieved or only the most relevant items or only unique/novel documents. There are different document types: they may be web pages, images,

videos, patents or research papers; they may differ in format (e.g., XML, PDF, HTML, OCR output, plain text) and in length (e.g., news articles, conference papers, abstracts, snippets, XML elements). There are different search environments: the set of documents over which the search ranges may be fixed or dynamic, large or small, flat or hierarchically structured; the environment may allow interactive search (where the user can modify their query based on initial retrieval results) or not. Thus, in experimental terms, there are too many variables in operational tests to be able to control enough of them so that observed results can be ascribed to definite causes.

Consequently, most evaluations take the form of laboratory tests. A laboratory test is a simulation of some real retrieval situation conducted in an artificial environment, rather than a test conducted in the real retrieval environment itself. In laboratory tests, the experimenter aims to improve real retrieval by modelling the real world and creating improvements in the modelled world. The more realistic the model, the more likely it is that observed improvements will also be observed in the real world. It is impossible to model the whole world, due to the very many parameters in real searches. Therefore, any laboratory test must limit what is modelled; the experimenter must decide what part of the world they are trying to improve and model that part. Thus, the laboratory model is not perfect; it is a necessary abstraction. However, the experimenter has control over all variables in their restricted, model environment; they can keep all factors constant except the independent variable they wish to investigate. In operational tests, the experimenter does not have control over these variables. Laboratory tests, therefore, offer increased diagnostic power over operational tests.

The Cranfield tests (Cleverdon 1960, Cleverdon, Mills & Keen 1966) are generally credited as the first laboratory tests and, furthermore, as having introduced the *test collection paradigm*. In this framework, a test collection comprises a document collection, a set of queries and, for each query, judgements of which documents are relevant to that query. Each of these three constituent parts is an abstraction of some part of the real world and, thus, has limitations and issues associated with it; we will discuss each of them later in this chapter. To evaluate a system, the experimenter submits the test collection queries to the system and the system retrieves documents; system performance is then evaluated by the relative positions of the relevant versus irrelevant documents in the retrieved document rankings.

One desirable property of a test collection is that it be *reusable*. A test collection is reusable if it is suitable to be used for the evaluation of different experiments, rather than being engineered solely for one specific evaluation conducted at one particular time. Reusable test collections are extremely valuable for several reasons. Firstly, test collections are very expensive to create. The cost of having to create a new test collection can prohibit IR experimentation. Secondly, as a fixed, static abstraction of the real world, the test collection allows for comparative evaluations between different systems, techniques and/or versions of systems at different times. The more evaluations a test collection can be used for, the more of a contribution it can make to the ultimate aim of improving retrieval in the real world.

Spärck Jones & van Rijsbergen (1976) discuss in detail the characteristics of the 'ideal' test collection. Such a test collection would be reusable in the extreme: it would be suitable for *every* conceivable retrieval experiment. Thus, the results of every retrieval experiment would be directly comparable. However, they concede that it is impossible for any one collection to have all of the characteristics they list: many of the requirements are contradictory. For instance, the ideal test collection would be heterogeneous in content (with documents from a range of subjects, varying in specialisation and hardness) *and* homogeneous in content. This is to allow for experiments where, e.g., subject matter is the independent variable under investigation as

well as experiments where it is not. Furthermore, test collections should represent real retrieval environments and these can be homogeneous or heterogeneous in subject matter. Spärck Jones & van Rijsbergen instead recommend the creation of a set of related, hierarchical collections such that experimenters can use different subsets or supersets of the collections that have certain common properties and certain distinct ones, depending on their particular experiment. However, such a set of collections has never been built.

Each part of a test collection is an abstraction of some part of the real retrieval world; we now consider the limitations and issues associated with each.

3.1.1 Queries

Queries as an abstraction of information needs

A test collection's query set is an abstraction of the innumerable user searches that are and will be carried out in the real world. There are three notable limitations with this abstraction: firstly, the fact that queries, not *information needs*, model user searches; secondly, the staticness of the queries and, thirdly, the finiteness of the query set. We consider each of these in turn.

The first limitation concerns the nature of searching. Real searches stem from an underlying *information need*; this is the gap in the searcher's knowledge that they wish to fill. To model a real search, what the experimenter ideally wants is an information need. However, information need is an intangible quantity, internal to the person with the need. To try to satisfy their information need, a searcher must express their need as a query, i.e., a representation of the need, bound by the constraints of the search environment. Taylor (1968) describes a continuous spectrum of question formation, from the unexpressed information need (the *visceral need*) through to the question as presented to the information system (the *compromised need*). Taylor's discussion is in the specific context of question negotiations between searchers in reference libraries and reference librarians, acting as the human intermediary in their search. Nevertheless, much of what is said applies to searches in general. In particular, there is an irresolvable difference between information need and what is eventually submitted to a retrieval system. Even in the real world, therefore, queries are a necessary, compromissary abstraction of users' information needs. This creates an upper bound for the experimenter hoping to model real searches in the laboratory; queries, not information needs, are the starting point for modelling user searches.

A second limitation of query sets is that the dynamic, adaptive nature of search is ignored and searches are modelled as single, stand-alone queries. This is generally not realistic; Taylor describes a user's search as 'merely a micro-event in a shifting non-linear adaptive mechanism'. In other words, what the user wants, expects and will accept as a satisfactory search result will change throughout the search process. A query is simply one expression of their need as they understand it at a particular point during their search. In modern retrieval systems, there is generally some element of interactivity in the search process; because results are usually returned quickly, users can browse the results from their initial query and submit a modified query based on the what has been retrieved. Teevan, Alvarado, Ackerman & Karger (2004) describe how users often choose not to specify their full information need straight away, even when they have a known target in mind. Instead, users conduct an *orienteering search* where they navigate towards their target in stages, gradually revealing their full need in successively more specific queries. This was supported in a study of searches on a digital library of technical reports, which observed that more than half of queries that followed a user's initial query built on the previous query (Jones, Cunningham & McNab 1998).

The third limitation is that of the finiteness of the query set. There are infinitely many

potential searches, from many different user types. Nevertheless, it is a practical necessity to model them with a finite set. To create as realistic a model as possible, though, the experimenter wants a realistic sample of searches, i.e., many different searches and search types, from many different users and user types.

Having noted the general limitations of the test collection query set, we now describe a number of ways in which a set of queries can be arrived at.

Procuring a set of queries

The ways in which queries can be procured can be split into two broad categories, which we will refer to as *manufacturing* and *observing*. In manufacturing, queries are artificially created for a given document collection. There is then the question of who creates those queries. The obvious easy option is for the experimenter to create the queries themselves. However, this will not satisfy the 'many users, many user types' model. Alternatively, some other person or persons might be recruited to create queries. For instance, a representative set of users might be hired to specify a number of queries they would typically search the document collection with.

An advantage of such manufactured queries is that the queries should be closely related to the documents; a test collection where the queries and documents are entirely disparate and do not have many judged relevant documents will not be useful as a comparative evaluation tool. Additional measures may be taken to more strictly control the degree of relatedness, e.g., an initial set of query suggestions might be filtered based on how many relevant documents are found in a preliminary search of the document collection. Manufacturing queries also has the advantage that the query creator is at hand to more fully specify the information need or, preferably, to make the relevance judgements for their query on the document collection. As we will see in Section 3.1.2, we assume that the query creator is the best judge of relevance with respect to their query.

On the negative side, manufacturing will be time-consuming and potentially expensive. Furthermore, manufactured queries may not necessarily stem from a genuine information need and, arguably, do not model real searches well. The test collection will be open to the criticism that the queries have been engineered (perhaps subconsciously) with a bias towards a particular experiment or system. This is particularly true of queries created directly by the experimenter.

Alternatively to manufacturing them, queries may be procured by observing users of retrieval systems and adopting their real queries for the test collection. This method has the advantage that the queries represent genuine information needs as formulated by real users. Observing queries may be one part of a larger study of human user behaviour. Saracevic, Kantor, Chamis & Trivison (1988), for example, observed users of a database system, studying the cognitive decisions and human interactions involved in the information seeking process, as well as the structure and classification of their search questions. Teevan et al. (2004) interviewed university computer scientists about their most recent email, file and Web search processes.

Rather than conducting a full user study, a cheaper way of observing real queries is to use query logs from operational retrieval systems. Logs provide a large snapshot of the queries that are occurring in the world; large enough to be used for statistical analyses. Queries taken from a log should give a representative sample of user types, query specificities etc. from the real world. A disadvantage of observing queries is that real queries are noisy: they can contain typographical errors, misspellings and ambiguities. Real retrieval systems should be robust to such imperfections. However, for many retrieval experiments, robustness will not be the primary focus and the effect of query noise may obscure the effects of the independent

variable that is being observed. Thus, it is often reasonable to 'clean up' noisy queries. A far bigger problem with query logs is that they provide limited insight into the information needs underlying the queries; without the user at hand and with limited search context, the experimenter cannot reconstruct the user's original need. This may detract from the value of having authentic queries.

In observational methods, there is also the question of which documents are in the observed retrieval system's index. If the system's documents are significantly different to those of the test collection, the queries taken from a query log may have little relation to the test collection documents. For instance, it is unlikely that many queries taken from a general-purpose Web search engine's logs will have any relevant documents in a collection of legal patents.

3.1.2 Relevance judgements

I can't get no satisfaction.
– M. Jagger and K. Richards

Relevance as an abstraction of user satisfaction

The fundamental idea in IR evaluation is quite simple: given a user's query, a system should return the documents that satisfy the user's information need and no others. We have already established that information need is a difficult concept. Just as it is non-trivial to define what a user's information need is, it is likewise difficult to pinpoint exactly what it means for that need to be satisfied; satisfaction is something internal, personal and mysterious even to the user, and will depend on the precise nature of their need. Unsurprisingly, different user types may be satisfied by different types of documents (or parts of documents) and by different numbers of documents. Consider, for example, a patent lawyer searching for existing patents that might conflict with a proposed new patent; the lawyer will only be satisfied by seeing *every* patent that is similar to the new patent idea so that they can determine that the idea is sufficiently original to be patented. On the other hand, a student searching for a definition of an unknown term may be satisfied with a single accurate snippet of a document. Less obviously, personal differences between users of the same type will also lead to differences in satisfaction, e.g., two students with that same definition search task may be satisfied by different documents. In some cases, a user might not even realise when their need has been satisfied. It might also be the case that their need simply cannot be satisfied. Satisfaction can be 'judged' only by the person with the information need, at the time of their need.

The term *relevance* is often used interchangeably with this difficult notion of user satisfaction. It is such a problematic quantity that the literature on the subject is extensive and periodically undergoes review (e.g., Saracevic 1975, Schamber, Eisenberg & Nilan 1990, Mizzaro 1997). What is considered relevant to a given query differs between people. What a given person considers to be relevant can differ over time and, also, can depend on what other documents they might have seen already. Spärck Jones (1990) sums this up by saying 'relevance is situational to a unique occasion'.

In the test collection paradigm, a simplified definition of relevance is used to model user satisfaction, where relevance means something like *topical similarity*, i.e., being about the same topic. Given a user's query, a system should return only the documents that are about the same

topic as the query. This makes relevance a relation between query and individual documents and has two notable consequences for IR evaluation: firstly, document relevance is independent of other documents; secondly, relevance is constant.

Defining relevance as topical similarity means that documents are relevant or irrelevant independently of each other: whether or not a document is about a topic is not affected by what other documents are about the topic. This is not a realistic model of relevance. Salton (1992) notes that the effectiveness of retrieval can be measured by determining the *utility* of documents to users and that utility and relevance are not the same: a document that is relevant to a query but does not contribute any information that is *novel* to the user is unlikely to be useful to them. In addition, Robertson (1977) notes that some documents may be relevant only when seen together, e.g., if they present complementary aspects of a topic. Nevertheless, it follows from the topical similarity definition of relevance that novelty of document content is unimportant and, hence, that the order in which relevant documents are retrieved and presented to the user is unimportant too.

A second, related consequence of the simplified definition is that relevance to a given query is an intrinsic, constant property of a document. Whether a document is about a topic or not will not change over time. Therefore, relevance can be defined by a static set of judgements per query as to which documents in the document collection are relevant to that query. Since relevance is intended to model user satisfaction, we make the basic assumption that the user (the query creator) is the person best qualified to make these relevance judgements.

A third simplifying assumption is that relevance is dichotomous: a document is either about a topic or not; a document is either relevant or irrelevant. This does not strictly follow from the topical similarity definition of relevance, since documents can be deal with several topics and may be more 'about' some topics than others; correspondingly, some documents may be more 'about' a given topic than other documents. Nevertheless, it is usual for the relevance judgements in a test collection to be binary. Again, this is different from real relevance, which is a more continuous phenomenon. For instance, some documents may satisfy a particular information need entirely, whereas others may only be helpful for identifying other relevant documents. Human judges generally report greater confidence when allowed to make relevance judgements along a multi-valued scale (Katter 1968). Some experimenters try to model something closer to real relevance by asking judges to grade documents for relevance according to some scale but these scales tend to be idiosyncratic and incomparable.

The model of independent document relevance is a major simplifying assumption and one that is known not to be valid in general. This assumption, however, allows the probability ranking principle (PRP) to be proven to hold (Robertson 1977). The PRP says that if a retrieval system ranks documents by decreasing probability of relevance to the user, it will be the most effective it can be to its users (Cooper 1972, cited after Robertson 1977). The PRP originated from Maron & Kuhns's (1960) idea that, since no IR system can be expected to predict *with certainty* which documents a searcher will find useful, a system must necessarily deal with probabilities. An important consequence of the principle's assumptions is that a system can produce a ranking by considering documents individually, rather than considering all possible rankings. This considerably reduces the computational complexity of the retrieval problem; Stirling (1975) investigated an algorithm to find the optimal document ranking but found it to be too computationally expensive. Moreover, an optimal ranking need not exist: Robertson (1977) discusses examples that demonstrate this.

Robertson concludes that the PRP is a general theory for retrieval in the case that documents

are relevant independently of each other but that there is no comparable theory that takes dependency into account. Schamber et al. (1990) called for an appropriately dynamic, situational definition of relevance to be developed but, still today, there is none. However, there has been experimental work in this area relatively recently. Thomas & Hawking (2006) introduce a tool for comparing whole result sets (not ranked sets, specifically), asking the user for their judgement as to which set they prefer for a given query. In this way, the question of whether individual documents are relevant and, therefore, whether their relevance is independent, is bypassed. This offers a potential solution to the inherent artificiality in after-the-fact judgements; it replaces the user's usual search interface and records their queries, interactions and judgements at the time of their search. Thus, the tool enables queries and real-time judgements to be gathered for real searches. However, because the tool elicits preference judgements between result sets and, specifically, between pairs of result sets, it could not easily be used for large-scale comparative, reproducible evaluations in the same way that test collections are.

Zhai, Cohen & Lafferty (2003) develop evaluation metrics for *subtopic retrieval* that reward systems for finding documents that cover as many subtopics within a query as possible and penalise systems for including redundant documents that cover the same subtopics. However, these methods are still very new and it remains to be seen whether they will become the basis for a viable theory of retrieval with interdependent document relevance. Thus, at present, the independent relevance judgements in a test collection may be regarded as a necessary, as well as convenient, abstraction.

As with queries, the methodologies by which sets of relevance judgements are produced can vary, as we will see in the following section.

Procuring relevance judgements

One of the most important factors in procuring relevance judgements is who the judge is. As for queries, there are several possibilities. The absolutely preferred option is for the query creator to make all of the relevance judgements; this is a more realistic model of user searches, where the person who issues a query to a retrieval system then makes some sort of evaluation of the quality of the documents that are returned to them. Alternatively, someone other than the query creator may make the judgements. In this case, the judge's familiarity with the query subject matter is an additional factor; whether an expert judge is more appropriate may depend on what sort of retrieval is being modelled. Finally, the task of making the judgements for a query may be distributed among some set of people. This is not realistic as a model of a single user's search. Judgements have been shown to differ between judges (e.g., Voorhees 1998) and, although Voorhees showed that the measured relative effectiveness of retrieval systems remained stable despite differences between judgement sets from different judges, these judgement sets were created by each judge making judgements for the same document set; it is not clear that distributing the judgements for the document set across the judges would likewise produce stable system rankings.

There is the further question of when the judgements are made. It is preferable that all the judgements be made (by the same person) at the same time, since a person's judgements have been shown to differ over time (e.g., Voorhees 1998). In the ideal case, the query creator would make relevance judgements straight after writing their query.

Given the simplified model of relevance, the way in which a document's relevance is judged will be unintuitive; the judges must be tutored in how to make their judgements within the constraints of the model, e.g., to judge documents independently of each other. Such tutoring

is usually done through written guidelines, telling them what constitutes a relevant versus irrelevant document. Nevertheless, the judge *will* make their judgements in some order; there is a danger that their perception of relevance may be altered by earlier documents they have judged, despite the guidelines (e.g., in those cases where document relevance is interdependent described by Robertson (1977)). Thus, the eventual set of judgements may differ depending on the order in which the documents are presented to the judge.

To make a test collection reusable, the relevance judgements should be *complete*, i.e., every document in the collection should be judged with respect to every query. This allows precision and recall, described in Section 6.2, to be calculated for any potential ranking of documents. Precision is the proportion of *retrieved* documents that are relevant. Thus, without complete judgements, precision can still be calculated, as long as the retrieved documents have been judged, i.e., those in the ranking considered in the calculation. Otherwise, precision for a new system can be calculated relatively cheaply, by making fresh judgements on the (potentially very few) top ranked documents. On the other hand, recall is the proportion of the *total* relevant documents that are retrieved. Therefore, recall calculations are particularly affected by incomplete judgements, since there is no way to accurately calculate the total number of relevant documents without judging every document.

The large scale of modern document collections makes it impossible to obtain complete judgements, however. The number of hours required to make complete judgements is far too high: Voorhees (2002) assumes a judgement rate of one document per 30 seconds and calculates that it would take over nine months to make complete judgements for a single TREC topic in an average TREC collection, which contains 800,000 documents. (See Section 3.2 for TREC.) Therefore, it is usually necessary to limit how many of the documents are judged. In particular, it is usual to try to find all documents that are relevant to a query, judge these and assume that unjudged documents are irrelevant. If every relevant document had indeed been found and then judged, this would be equivalent to having complete judgements. In the *pooling method* (Spärck Jones & van Rijsbergen 1976), used in TREC, the top documents from some number of retrieval systems or methods are merged and only the documents in this pool are judged. The aim is to find all relevant documents in the pool, while keeping the pool as shallow as possible. To this end, as diverse as possible a range of systems should contribute to the pool; it then becomes less likely that any future method (that did not contribute to the pool) will return a relevant document that none of the pool methods has returned. Both automatic systems and manual searches may contribute to the pool. The more unique relevant documents a method contributes to the pool, the more valuable that method is for the aim of finding all relevant documents.

With the scale of modern test collections, it becomes increasingly infeasible to be confident that pooling will return all relevant documents. Zobel (1998) estimates that at most 50-70% of relevant documents have been judged in some TREC test collections. However, despite this incompleteness, Zobel found that a pool depth of 100 (the usual depth of TREC pools) is sufficient to give reliable evaluation results, both for pool systems and other systems, i.e., measurements of relative system effectiveness are trustworthy and fair. Even so, pooling requires a huge judgement effort and associated expense. Therefore, there has been various work on methods for selecting smaller sets of documents for judgement from the pool while trying to maintain reliability.

Zobel (1998) proposed a variation of pooling where the pool depth is gradually incremented per query. Whether a query's pool is deepened and judged at each stage depends on how likely

it is that any of its new pool documents will be relevant, according to regression on the number of new relevant documents that were found at previous pool depths. Thus, the judgement effort is concentrated on those queries that are likely to have more relevant documents; more relevant documents should be found for a given amount of effort and the reliability of the measured results should be increased. More recently, Aslam, Pavlu & Yilmaz (2006) use a technique based on random sampling to select documents from the pool for judging, rather than judge the entire pool. Their method allows an accurate estimation of the values of the standard evaluation measures that would have been calculated using the full TREC pool, with only a small fraction of the judgements. Carterette, Allan & Sitaraman (2006) present an algorithm for selecting the document from the pool that will maximise the discriminatory power of the set of judged documents, i.e., maximise the difference in average precision between systems contributing to the pool. Their method is interactive: after each selected document is judged, its relevance judgement is taken into account to calculate the difference in average precision that each remaining unjudged document will create, if it is judged. The theory behind the algorithm suggests a natural stopping condition that indicates when sufficient documents have been judged, i.e., sufficient to prove that the contributing systems are different in terms of average precision. Thus, redundant judgement effort is avoided.

3.1.3 Document collection

Documents as an abstraction of search space

The documents in a test collection model the search space of a real retrieval environment. Searches in the real world are conducted over a wide variety of search spaces, differing in size, document genre, document length, document type and heterogeneity of documents in all of these respects. The search space need not necessarily be static. As an extreme example, the Web is a highly dynamic collection of mostly HTML documents; Ntoulas, Cho & Olston (2004) estimated that, every week, 8% of existing web pages are replaced with around 320 million new pages and 25% new hyperlinks are created. Other, less apparently dynamic document collections may grow gradually over time, e.g., as new generations of documents are added to archives. In the laboratory model, this is abstracted over. The document collection used in this model is necessarily static: complete relevance judgements would be impossible on an unknown set of documents. Issues arising from the changeable nature of dynamic collections, therefore, cannot be easily investigated using test collections.

What makes a suitable document collection will depend on what retrieval situation is being simulated, i.e., a collection of news articles is clearly unsuitable for evaluating techniques for retrieving chemistry papers. Certain types of retrieval naturally produce queries with an element of currency or time-dependency, e.g., searching for news on current affairs. In such cases, the documents must be synchronous with the queries. This is a special case of the notion of query-document relatedness introduced earlier.

Similarly, the nature of the particular experiment or evaluation may impose requirements on the document collection. An experimenter investigating the relative effects of different techniques for document length normalisation, for instance, will need a collection with documents of varied length. An evaluation of system performance on very specific queries (i.e., with few relevant documents) will require a collection with high density of similar documents, to distinguish between systems that can and cannot identify relevant documents from similar irrelevant ones.

A further issue is the size of the document collection. Many real searches are conducted over

very large document collections so, for a realistic model, a large collection is often desirable. It is generally accepted that the more documents, the better. However, the question of whether there is a minimum 'good' number is, as yet, unaddressed; in fact, the statistical effect of such large numbers of documents on evaluation measure reliability is currently unclear (Robertson 2007).

Procuring a document collection

Having determined the appropriate characteristics of the document collection, there are two broad categories of ways in which a collection may be procured: we will distinguish between *manufactured* and *real* document collections. Firstly, a suitable set of real documents may be obtained from various sources and artificially grouped together by the experimenter. So-called manufactured document collections are subject to similar criticisms as manufactured queries: they may be engineered with a certain experiment or system in mind, leading to a bias in any evaluation results. Also, since they are manufactured, it could be argued that they do not model real document collections well; real collections are cohesive sets of documents that have been grouped together for some real purpose, independently of any IR experiment.

Alternatively, a real document collection may be appropriated. This has the advantage of authenticity; results found on real collections are more likely to carry over to the real world. Using a real collection has the particular advantage that results found in the laboratory may be directly applied to the same collection in the real world. As noted, dynamic collections cannot be used in a test collection so using a real collection is not an option in, e.g., Web retrieval. However, a static set of Web documents may be created in a number of ways, some of which are more realistic than others. For instance, the results of a Web crawl performed provide a 'snapshot' of the Web at a particular time. On the other hand, pages from a certain limited domain may be used.

In practice, the choice of documents for a test collection will be largely determined by practical considerations, e.g., copyright issues, what documents (or document collections) are available and in what format, whether a financial cost is associated with certain documents, how large a collection can practically be managed etc. It is usually the case that some degree of compromise will be necessary when procuring a document collection.

3.2 Methodologies for building test collections

Although a universally useful, 'ideal' test collection does not exist, there are a number of test collections available for IR experimenters. In this section, we examine a number of notable test collections and how they were built. TREC is a large-scale, government-funded IR conference which has produced several test collections for different IR tasks, e.g., on news text (the Ad hoc task), Web material (the VLC/Web track) and biomedical documents (the Genomics track). INEX is an initiative specialising in XML retrieval, which has adapted the TREC methodology to create test collections for its own requirements. Finally, Cranfield 2 is a one-off comparative evaluation of indexing language devices, which produced a small test collection of scientific paper abstracts.

3.2.1 TREC

The Text REtrieval Conference, run by NIST[1] and ubiquitously known as TREC, opened large-scale IR evaluation to the research community in 1992 (Voorhees & Harman 2005). In 2006,

[1] National Institute of Standards and Technology, United States Department of Commerce

there were 107 participating groups; the year before, there were 117. TREC evolved from the DARPA[2] TIPSTER project (Merchant 1994, Altomari & Currier 1996, Gee 1999), whose aim was to improve IR and data extraction from real, large data collections.

The main TREC task was the Ad hoc task, which ran for eight years. In TREC-4, the conference expanded to include tracks, i.e., secondary tasks that focused on particular aspects of the original tasks, e.g., the Robust track used Ad hoc queries that were selected precisely for being 'hard', or introduced new areas of retrieval research, e.g., the Video, SPAM, Question Answering and Genomics tracks. Several specialised test collections were built for the purposes of these tracks. The construction of these collections has roughly followed the methodology for the Ad hoc collections, with task-dependent deviations as necessary. The TREC test collections are widely used as the data for current retrieval research (e.g., Carterette 2007, Mizzaro & Robertson 2007, Zhou & Croft 2007).

Ad hoc

The (English) Ad hoc collections are the most prevalent and heavily used of the TREC collections. The Ad hoc retrieval task was devised to model searches by expert users who work intensively with information in large quantities and require high recall: information analysts, primarily. In each year of the task, the test collection was updated slightly, as new *topics* (i.e., TREC terminology for queries) were created and additional documents were procured. Harman (2005) gives a detailed, year-by-year account of how the collections were built and analyses the resultant collections.

Documents: The genre of the Ad hoc document collections has always been news (i.e., newspaper and newswire articles), patents and documents from various government departments. The documents were mostly donated but, for some, the usage rights had to be purchased. The majority of the documents are copyrighted so permission to use and distribute those documents also had to be obtained. The exact set of documents used each year varied slightly but always included over half a million documents, to model the large corpora that information analysts typically search.

Queries: In each TREC, exactly 50 topics were used. Topics were created by information analysts hired by NIST, i.e., real users from the type of retrieval being simulated. The topic authors were instructed to devise topics with the specific documents for that year in mind. Trial searches were conducted with these topic ideas and ones for which roughly 25 to 100 relevant documents were found were selected as a topic. The intention behind this filtering was to ensure that the topics represented a range of broader to narrower user searches. Over the years, the guidelines for topic creation were gradually adjusted to try to ensure that the topics were as realistic as possible: originally, topic ideas were inspired directly by the documents and could be modified later in the creation process; later, the topic creators were instructed to use their own genuine information needs and their topics were not modified.

The guidelines also specified the structure of topics, which comprised a number of named fields each designated to specify some aspect of an information need. For instance, the narrative field describes what constitutes a relevant document to that topic. The detailed topic design was intended to provide a more thorough representation of an information need, as opposed to more traditional queries. This was to allow for broader research (i.e., into query construction methods), to make it easier for the relevance judges to make consistent, independent judgements

[2]Defense Advanced Research Projects Agency, United States Department of Defense

and, also, to make the topics more understandable and, thereby, make the collections more reusable.

Relevance Judgements: The relevance judgements were also made by hired analysts, always by a single person for a given topic (for consistency) and, except in TREC-1 and TREC-2, by the topic author (to emulate a real search). Pooling was used; each of the systems participating in the task contributed its top 100 documents to the pool. The judges were instructed to judge a document as relevant if it contained any information that would be useful in writing a report on the topic, to model the high-recall searches of information analysts.

VLC and Web

The Very Large Collection (VLC) and Web tracks were intended to provide a test bed for studying issues from retrieval on document collections orders of magnitude larger than the Ad hoc collections. See Hawking & Craswell (2005) for a thorough account of both tracks and their associated collections. The VLC track came first and was devised to investigate retrieval efficiency and scalability. The focus shifted towards evaluating Web search tasks and the VLC track was naturally succeeded by the Web track. As the track focus diverged from the Ad hoc model, so too did the methods used in constructing the test collections. We highlight some of the differences here.

Documents: The first VLC track simply modelled the same type of retrieval as the Ad hoc track but on a larger scale. Thus, the document collection was very similar to the Ad hoc collections in genre but much larger: 7,492,048 mainly news and government documents, with a small proportion of Web text. This was replaced in TREC-7 with VLC2, the results of a Web crawl from the Internet Archive[3], totalling 18,571,671 documents.

Three collections for evaluating Web search tasks were created: WT2g, WT10g and .GOV. These collections could be smaller, since the focus was not on issues stemming from the size of the Web. Two were created by selecting subsets of VLC2 to suit a specific task definition, i.e., to investigate whether hyperlinks could be used to improve ad hoc retrieval on web pages. WT2g is the smallest subset, created to allow as complete relevance judgements to be made as in the Ad hoc collections so that more direct comparisons could be made with Ad hoc track results; WT10g is a larger subset specially selected to ensure a high proportion of inter-server links. Thus, these document collections were artificially created from the more authentic Web crawl collection to have characteristics suitable for specific experiments. These were succeeded by the .GOV collection, created by a truncated crawl of the .gov domain, totalling 1,247,753 documents. This collection is more manageable in size than the full Web collection but is still a natural, cohesive subset of the Web. Thus, the track has modified its methodology to create a document collection that is as suitable yet realistic a model as possible.

Relevance Judgements: Shallower pools than in Ad hoc (by a factor of five) were used for obtaining relevance judgements on the main collections (Harman 2005). The reasons for this were twofold. Firstly, because these collections are so much larger than any Ad hoc collection, it would be impossible to assume even sufficiently complete relevance judgements. Secondly, the tracks were designed to reflect Web searches, rather than the high-recall searches of information analysts, and the evaluation focus shifted to achieving early precision.

A related difference is in the use of graded relevance judgements in some tracks. In response to the argument that, when searching the Web, users are interested in finding some

[3] `http://www.archive.org`

highly relevant documents, rather than all relevant documents, three-level judgements were introduced in the main Web task, to explore the role of highly relevant documents in system evaluation (Voorhees 2001).

Queries: For the VLC track, each year's Ad hoc topics were used along with the current VLC document collection. When VLC was succeeded by various Web tracks, e.g., Web Topic Relevance, Topic Distillation etc., specialist topics were used instead, to better simulate those Web search tasks. For example, in the Large Web task, the 50 queries used were extracted from Web query logs. In later tracks, e.g., Web Topic Relevance, traditional TREC-format topics were reverse-engineered from query log queries. Thus, real queries were used but the underlying information needs had to be guessed. In the Home-page Finding task, a random selection of home pages within the collection were used as 'topics', i.e., targets.

Genomics

The Genomics track was created due to interest in experimenting with more structured types of data than newswire and, particularly, in using data in public databases. The genomics domain was chosen due to availability of resources. Initially, ad hoc retrieval was the main task; in 2006, the track focus shifted to passage retrieval for Question Answering. In each year of the track, as more resources became available and the track focus evolved, new test collections and methodologies were developed. See the track overviews for more details (e.g., Hersh & Bhupatiraju 2003, Hersh, Cohen, Roberts & Rekapilli 2006).

Documents: The original Genomics collection consisted of 525,938 MEDLINE records, which contain (at most) abstracts and never full documents. This was replaced in 2004 by a 10-year subset of MEDLINE, totalling 4,591,008 records. For the 2006 track, a new collection of 162,259 full-text biomedical articles was created from a Web crawl of the Highwire Press site, with the permission of the publishers who use that site to distribute their journals.

Queries: Topics for the 2003 ad hoc retrieval task were gene names taken from a public database. The task definition for that year was very specific: roughly, given a gene name, return all documents that focus on the basic biology of that gene. Gene names were randomly selected from across the spectrum of genes in the database, according to various criteria, e.g., the number of Gene Reference Into Function (GeneRIF) entries that gene has in the database. A GeneRIF is a statement about one particular function of a gene, paired with a MEDLINE reference to the article that discovered that data. In subsequent tracks, the topics were based on genuine information needs from biologists; free-form biomedical questions, initially, and, eventually, structured questions derived from a set of track-specific topic types, e.g., What is the role of *gene* in *disease*? In each year, there were 50 topics, except in 2006 when only 28 topics were used.

Relevance Judgements: For the gene name task in 2003, the MEDLINE references from GeneRIFs were used as pseudo-relevance judgements. The track had limited resources and this method is extremely cheap, since no relevance judges are required. However, a GeneRIF entry pairing an article with a gene name is *not* an explicit judgement of relevance, in terms of the task definition. Furthermore, the GeneRIF 'judgements' are known to be incomplete. In the later tracks, when more resources were available, pooling was used and explicit relevance judgements were made by biologists. In 2004, the average pool depth across queries was 75 documents and there were two judges; one PhD student and one undergraduate student in biology. In 2005, the pool depth was 60 and there were five judges with varying levels of expertise

in biology. In 2006, due to the Question Answering nature of the track, passages not full documents were judged; stricter guidelines for judging were produced and, additionally, judges were given one hour of training. There were nine 'expert' judges and 1000 passages were judged per topic.

3.2.2 INEX

The INitiative for the Evaluation of XML retrieval, or INEX, was founded with the aim of providing an infrastructure for content-oriented XML document retrieval, including a large test collection of real XML documents (Gövert & Kazai 2002). The main INEX task is ad hoc retrieval on XML documents, with additional XML-oriented tasks being added in later years. Like in TREC, the test collection has been augmented for each annual workshop and both the retrieval tasks under investigation and the methodologies used to build the test collection have evolved over time. See the workshop overviews for more detail on each year's changes (e.g., Gövert & Kazai 2002, Malik, Trotman, Lalmas & Fuhr 2006).

Documents: The original document collection consisted of 12,107 full-text IEEE Computer Society articles of varying length with XML mark-up. In 2005, 4712 new articles were added. In 2006, a new collection of 659,388 Wikipedia articles was introduced. Thus, INEX has two document collections; one with a fairly broad but restricted domain and one open-domain collection. Both collections were donated; the genres seem to have been determined by document availability rather than by the goal to model a specific retrieval environment.

Queries: The methodologies for queries and judgements closely followed those for TREC Ad hoc with a few notable deviations. Firstly, the TREC topic format was modified to incorporate XML structural conditions. More interestingly, queries and judgements were created by the groups participating in the workshop, following written INEX guidelines. This circumvents the expense of creating queries and judgements. It also affected the query selection process in two ways. Firstly, each participating group used their own system to retrieve the top documents used in the filtering stage of candidate queries, in the first year. Since 2003, a proprietary INEX retrieval system was used in the filtering stage for each candidate query, giving a more consistent treatment across queries. Secondly, the criteria for query selection were extended so that each group was allocated roughly the same number of queries, i.e., to evenly distribute the judgement effort across groups.

Judgements: The relevance judgements for each query were made by the group that contributed the query or, where that was not possible, by a group who volunteered with knowledge in the query subject area. Thus, there is less control over who the judge is than in TREC; the judgements may be made by a different member of the same group as the query author, by multiple different members of that group or by a member or members of a different group. The judgements made differ from traditional relevance judgements: graded judgements along two dimensions were made, called *exhaustivity* and *specificity*[4]. The definitions of these dimensions are specific to the nature of the XML task; we do not discuss them in detail here. However, in 2005, by its definition, specificity could be measured automatically by INEX's online assessment tool (according to the ratio of text within an XML component that was judged relevant by the judge), reducing the manual judgement effort.

[4]These dimensions were called *topical relevance* and *component coverage*, respectively, in INEX 2002.

3.2.3 Cranfield 2

Cranfield 2 was an independent comparative evaluation of indexing language devices. After World War 2, conventional indexing methods could not cope with the great quantity of scientific and technical reports newly released from security restrictions. Consequently, many new experimental indexing techniques were developed, causing arguments between the advocates of each technique, as well as with professional librarians, who defended their traditional methods. Cranfield 2 was intended to settle the controversy, while tackling some of the methodological criticisms of earlier tests (Cleverdon 1960). See the project report for a more complete description of the Cranfield 2 methodology (Cleverdon et al. 1966).

The principles behind the Cranfield 2 methodology are that every scientific paper has an underlying research question or questions and that these represent genuine information needs and, hence, valid search queries; that a paper's reference list is a source of documents relevant to its research questions and that the paper author is the owner of the information need and, therefore, the person best qualified to judge the relevance of their references. Thus, papers are a source of search queries, as well as a list of potentially relevant documents, where the query author is known; papers were the starting point for building the Cranfield 2 test collection.

The methodologies for creating the document collection, queries and relevance judgements are very much interdependent and less separable than in the test collections we have already looked at: the relevance judgements were interleaved with the query creation process; the final document collection was a product of this process. We will discuss the queries first.

Queries: The queries were created by the authors of a base set of scientific papers. The authors were invited to formulate the research question behind their work and, optionally, up to three additional questions that arose during that work, that they had (or might have) used as searches in an information service. The written guidelines specified that questions should be given as natural language sentences. Questions were selected that were grammatical and had sufficiently relevant references, as will be discussed later. A small number of questions were reformulated to remove anaphoric references, which were the effect of authors returning a series of interrelated research questions; all modifications were approved by the question author. The majority of the original query authors completed a second round of judgements, giving the final query set (Cleverdon 1997). Thus, the Cranfield 2 queries represent genuine information needs; the intention behind the methodology was to seek out the owner of the need to write the query. Furthermore, the source of these information needs, i.e., the papers, were also a source of potentially relevant documents: each paper's reference list was a list of documents that were probably relevant to at least one of its underlying questions. This was the starting point for making relevance judgements.

Judgements: In their invitation to participate, the authors were also asked to judge how relevant each reference in their paper was to each of the questions they had given. Relevance was measured on a 5-point scale, defined in terms of how useful a reference was in answering a given question, where grade 1 is most useful. Grammatical questions with at least two references judged as grade 1, 2 or 3 were selected for a second round of judgements. The reference list in scientific papers is generally not an exhaustive list of documents that are relevant to the issues in the paper; in a collection of documents from the same field, there are likely to be other relevant documents. In other words, judging the references alone is unlikely to give complete relevance judgements. The scale of the Cranfield 2 collection was small enough that trying to get complete judgements was feasible.

In the second round, a list of potentially relevant additional documents was sent to the author, with an invitation to judge the new documents for relevance, using the same relevance scale. The authors were also asked to weight the terms in their questions according to a 3-point scale of how important that term is to the question and to list any other search terms for the question or a complete reformulation of the question, if necessary. This was to allow for experimentation on the relative importance of query terms. Materials were sent to the authors to try to make the judging process easier, e.g., the author's original question(s) and the abstract of each of the new documents for judgement. Thus, all relevance judgements were made by the query author.

The list of potentially relevant documents was created by a combination of two methods. Firstly, the document collection (on paper) was searched by hand. The queries were grouped into small batches with very similar subject area and, for each batch, the entire collection was searched. Grouping the queries by subject was designed to reduce the search effort. The searching was done by post-graduate students with knowledge of the field. The students were instructed to list any document that they suspected might be relevant to a given query. The manual searches were designed to be thorough: every document in the collection should have been considered for relevance with respect to every query, by a non-novice in the field, who made liberal selections based on potential relevance. In addition to these manual searches, bibliographic coupling was used to retrieve documents similar to the already judged relevant documents. Thus, for each question, documents sharing seven or more references with the source document's judged relevant references were included in the list of potentially relevant documents. The intention was that the list presented to the final judge would contain every relevant document in the collection, rendering the final set of relevance judgements complete, effectively.

Documents: The Cranfield 2 collection started from a base set of scientific paper abstracts. These papers were mostly on high speed aerodynamics but also included some on aircraft structures, so that the effect of having documents on two dissimilar topics could be examined. The genre was determined by document availability, by virtue of the evaluation being conducted at Cranfield College of Aeronautics. The base documents were selected on the basis of being published recently, being written in English and having at least two English references that were published no earlier than 1954 and were likely to be easily obtained. The final document collection consists of a) the base documents (abstracts) for which authors returned research questions, b) their cited documents and c) around two hundred additional documents 'taken from similar sources'.

The Cranfield 2 test collection consists of 221 scientific queries and a manufactured collection of 1400 abstracts with a very limited domain. By design, the queries and documents are closely correlated, because the source document and cited documents for each query are automatically included in the collection.

Criticisms of Cranfield 2 methodology

Both Cranfield 2 (Cleverdon et al. 1966) and its predecessor Cranfield 1 (Cleverdon 1960) were subject to various criticisms; Spärck Jones (1981) gives an excellent account of the tests and their criticisms. The majority were criticisms of the test collection paradigm itself and are not pertinent here. However, the *source document principle* (i.e., the use of queries created from documents in the collection) attracted particular criticisms. The fundamental concern was that the way in which the queries were created led to 'an unnaturally close relation' between the query terms and those used to index the documents in the collection (Vickery 1967). In other

words, the author chose the words in the query after choosing the words in the source document when they wrote the paper. The words in the document will be used to index the document. Thus, the query terms are not inspired independently of the index terms. Any such relationship might have created a bias towards a particular indexing language, distorting the comparisons that were the goal of the project. A particular closeness between the query terms and source document index terms would also create a bias towards retrieving these documents above others.

In Cranfield 1, system success was measured by retrieval of those source documents and *only* those source documents. This was criticised for being an over-simplification and a distortion of real searching: in general, queries do not have a source document, nor even one pre-specified target document. Furthermore, the suspected relationship between queries and source documents would have a critical effect in this case. The evaluation procedure was changed for Cranfield 2 so that source documents were excluded from searches and, instead, retrieval of other relevant documents was used to measure success. This removed the main criticism of Cranfield 1. Despite this, Vickery notes that there were 'still verbal links between sought document and question' in the new method: each query author was asked to judge the relevance of the source document's references and 'the questions ... were formulated *after* the cited papers had been read and has [sic] possibly influenced the wording of his question'.

The source document principle, thus, carries a potential risk of artificially close links between queries and particular documents in a test collection.

3.3 ACL Anthology test collection

The particular kind of experiments proposed for our research imposes certain requirements on the test collection which rule out the use of pre-existing test collections. Firstly, we require documents with citations. This makes the TREC Ad hoc and VLC/Web test collections unsuitable. Secondly, we require the full text of both the citing and cited paper. This rules out the earlier TREC Genomics test collections and, for instance, the German Indexing and Retrieval Test (GIRT) collections (Kluck 2003), in which most documents are sets of content-bearing fields, not full-text documents. The CACM collection of titles and abstracts from the Communications of the ACM is likewise unsuitable. The 2006 TREC Genomics collection of full-text documents was not available when our work began and, regardless, contains judgements for a Question Answering task rather than document retrieval (Hersh et al. 2006). The original INEX collection of IEEE articles contains some citations but was only beginning to be made available as our test collection effort was already underway. Thus, there was no ready-made test collection that satisfied our requirements. Therefore, a substantial portion of the research effort of this thesis went into designing and building a test collection around an appropriate document collection, the ACL Anthology, to be discussed in Chapter 4.

We have seen from the previous section that there are alternative methods for constructing a test collection. The TREC Ad hoc methodology has become somewhat epitomic; the collections for later TREC tracks, as well as more recent comparative evaluations, such as INEX, have largely been built following the TREC model. This methodology is extremely expensive, however. Some of the TREC costs have been circumvented in later collections, e.g., the cost of hiring professionals to create queries and make relevance judgements is removed in INEX by requiring the workshop participants to contribute this data before they are granted access to the final collection. Nevertheless, that process is still extremely labour-intensive, requiring many collaborators to contribute many hours of effort.

The TREC methodology was an unrealistic option in the context of this research. The Cran-

field 2 methodology, on the other hand, is a far cheaper alternative and, since we aim to work with scientific papers, it was also readily applicable to our requirements and our resources. We, therefore, altered the Cranfield 2 design to fit an existing, independent document collection and applied it to the ACL Anthology. The following section documents that process and describes the resultant test collection. Appendix B.1 gives a detailed comparison between the Cranfield 2 methodology and our adapted methodology.

In Section 3.2.3, we described the source document principle and the associated criticisms. Though we have adapted the Cranfield 2 methodology, we too have source document queries and must consider these criticisms. Firstly, we discount retrieved source documents from our evaluation, in keeping with Cranfield 2. Secondly, our test collection is not intended for comparisons of indexing languages. Rather, we aim to compare the effect of adding extra index terms to a base indexing of the documents. The influence that the source documents will have on the base indexing of a document is no different from the influence of any other document in the collection. The additional index terms, coming from citations to that document, will generally be 'chosen' by someone other than the query author, with no knowledge of the query terms[5]. Also, our documents will be indexed fully automatically, further diminishing the scope of any subconscious human influence. Thus, we believe that the suspect relationship between queries and indexing is negligible in the context of our work, as opposed to the Cranfield tests. There is also, however, the question of whether citation terms from the source documents should be excluded from our indexing; if the source document terms do give an unnatural advantage when used as index terms, our method passes this advantage on to the source document's cited papers. This is an open issue that we do not investigate in this thesis but reserve for future work.

3.3.1 Pilot study

While designing our methodology, we conducted a pilot study on members of the Natural Language and Information Processing research group and some of its close associates. This was done in two stages. Firstly, informal one-on-one interviews were held with a small number of people, discussing in detail their reasons for citing each paper in the reference list of one of their papers. The aim of these discussions was to gain some insight into what 'types' of reference there are and how they vary in importance or relevance. We drafted a relevance scale based on the findings, trying to make the relevance grades and their descriptions reflect how paper authors judge their references. We defined a new, 4-point relevance scale, since we felt that the distinctions between the five Cranfield 2 grades were not appropriate for the computational linguistics domain and, also, we hoped to make the judgement task easier for our paper authors. We chose not to simplify to binary judgements, despite the fact that the standard evaluation measures assume binary relevance. Firstly, there is evidence that the extra information in graded judgements is useful in distinguishing between systems' performance (Järvelin & Kekäläinen 2002); asking for binary judgements would rule out possible experimentation, since binary judgements cannot later be expanded into more distinct categories. Conversely, it is possible to convert graded judgements to binary: graded judgements have been collapsed in previous studies and shown to give stable evaluation results (Voorhees 1998).

The possibility of a second round of judgements complicated the decision to ask for graded

[5]Self-citation is the exceptional case. This would in theory allow the query author to influence the indexing. However, it seems highly improbable that an author would be thinking about their query whilst citing the source document as previous work. Moreover, it would require malicious intent and a knowledge of our methods to have an adverse impact on our experiments.

judgements. The relevance scale used in Phase One was designed for the specific task of grading the relevance of cited papers in relation to the research question underlying the source paper; the grades were described in terms of how important it would be for someone reading the paper to read that cited paper. Judging the relevance of other papers (i.e., papers not cited) is a slightly different task and would have required a translation of the relevance scale. It was not clear that a directly interchangeable set of grades could have been formulated, such that a Phase One grade 4 was equivalent to a Phase Two grade 4 etc. Nevertheless, it was not clear when planning Phase One whether Phase Two would definitely be carried out and we opted to use graded judgements for Phase One.

The second stage of the pilot study took the form of a trial run of the data collection. Invitations to participate and the pilot materials were emailed to around a dozen subjects from the same group of people as in the first stage. We tried to simulate the real invitation scenario as closely as possible: the invitees were given no prior warning or priming and we used their most recent conference paper. As well as the altered relevance scale, we extended the Cranfield 2 design to invite authors to list any relevant documents that they knew of from outside their reference list, to try to increase the number of judged relevant documents. Three subjects participated and were then prompted for feedback, e.g., how difficult they found the task, how long they spent on it, whether they found any parts of the task particularly difficult. Another invited subject gave useful feedback without participating.

We received no feedback from the pilot study to suggest that our four grades were not expressive enough. The major design change that came out of this stage was that we decided against asking authors to give additional relevant documents. One author found this to be a difficult, potentially open-ended task. Also, we realised that an author's willingness to name such documents will differ more from author to author than will their choosing the original references. This is because referencing is part of a standardised writing process that they have already completed, whereas optionally naming other relevant documents would take up more of their time. By asking for this data, the consistency of the relevance data across papers would be degraded and the status of any additional judgements would be unclear. Since additional relevant documents would, in principle, be identified in a second round of relevance judgements, not asking for additional relevant documents in the first round should not result in relevant documents being missed.

3.3.2 Phase One

We altered the Cranfield 2 design to fit to an existing document collection, rather than creating an artificial collection from the source documents and their references. This gives us all the advantages of using a real document collection, rather than a manufactured one; we are modelling retrieval in a realistic environment. We designed our methodology around an upcoming conference that would be archived in the ACL Anthology (ACL-2005). We have several motivations for this aspect of the design. Firstly, we know that, in any single conference year, there will be a large number of papers presented, from which we might get queries. Secondly, we assume that, in any single conference year, papers from across the field will be presented, thereby giving us a range of queries from the domain of the document collection. Thirdly, we assume that, in any single conference year, there will be papers by many different authors. Taken together, these factors help us create a 'many queries, many users' model. Fourth, we assume that authors from the conferences that are archived in the Anthology are people who are likely to search the Anthology. As in Cranfield 2, our queries should represent genuine information needs; we

further assume that conference papers (as opposed to journal papers) are written fairly recently in advance of the publication so the information need is, likewise, fairly recent. This serves to minimise the time between the original information need and the author formulating their query for us. In summary, by asking for queries from recent conference authors, we hope to procure a set of queries that is a realistic model of searches that representative users of the document collection would make.

We approached the paper authors at around the time of the conference, to maximise their willingness to participate and to minimise possible changes in their perception of relevance since they wrote the paper. Due to the relatively high *in-factor* of the Anthology, we expected a significant proportion of the relevance judgements gathered in this way to be for other Anthology documents and, thus, useful as test collection data. (See Chapter 4 for in-factor and its bearing on the test collection.) We applied our methodology to two separate Anthology conferences, to try to gather as many queries as possible; ACL-2005 in May 2005 and HLT-EMNLP-2005 in October 2005.

The authors of accepted papers were asked, by email, for their research questions and for relevance judgements for their references. A sample email and other materials are reproduced in Appendix B.2 and our relevance scale is given in Table 3.3. Personalised materials for participation were sent, including a reproduction of their paper's reference list in their response form. This meant that invitations could only be sent once the paper had been made available online, either on the author's own web site or through the Anthology web site.

Each co-author of the papers was invited individually to participate, rather than inviting only the first author. This increased the number of invitations that needed to be prepared and sent (by a factor of around 2.5) but also increased the likelihood of getting a return for a given paper. Furthermore, data from multiple co-authors of the same paper can, in principle, be used to measure co-author agreement on the relevance task. This is an interesting research question, as it is not at all clear how much even close collaborators would agree on relevance; our methodology should allow for investigation of this issue.

3.3.3 Phase Two

In line with the Cranfield 2 methodology, we expanded our test collection in a second stage. The returns from Phase One are summarised in Section 3.3.4. We conducted some analytical experiments with this data and observed that the values of, e.g., MAP, R-precision and P@5 increased when queries with lower than a threshold number of judged relevant documents were excluded from the evaluation (Ritchie, Teufel & Robertson 2006*a*). (See Section 6.2 for definitions of these evaluation measures.) Based on these results, we decided that the relevance judgements at this stage were too incomplete and that a second round of judgements was necessary, though the Anthology is too large to be able to expect complete judgements. The purpose of our Phase Two was solely to obtain more relevance judgements for our queries, to try to bridge the completeness gap. We used the pooling method to identify potentially relevant documents for each of our queries, for the query authors to judge.

We reformulated some of the research questions returned in Phase One. Upon studying the questions, we identified several ways in which a number of them were unsuitable as queries. Mostly, these were artefacts of the method by which the queries were created: we did not explicitly ask the authors for *independent* search queries. Hence, where an author had returned multiple research questions, the later questions sometimes contained anaphoric references to earlier ones or did not include terms describing the background context of the research that had

Reformulation	Description
Typo	Corrected spelling or typographical error in the research question, as returned by the author.
Filler	Removed part(s) of the research question that did not contribute to its meaning, e.g., contentless 'filler' phrases or repetitions of existing content.
Anaphor	Resolved anaphoric references in the research question to ideas introduced in earlier research questions from the same author.
Context	Added terms from earlier research questions to provide apparently missing context.

Table 3.1: Reasons for query reformulations.

been introduced in an earlier question. In addition, some questions contained spelling or typographical errors and some were formulated elaborately or verbosely, with many terms that did not contribute to the underlying meaning, e.g., contentless rhetorical phrases or repetitions of existing content. Robustness to such query imperfections is outside the domain of our research. Therefore, in line with Cranfield 2, we minimally reformulated 35 of the 201 research questions into error-free, stand-alone queries, while keeping them as close to the author's original research question as possible. Table 3.1 describes the four classes of query reformulation; Appendix B.3 gives a complete list of the reformulations we made.

For each query, we next constructed a list of potentially relevant documents in the Anthology. The present author first 'manually' searched the entire Anthology using the Google Search facility on the Anthology web site, starting with the the author's complete research question (or our reformulation) as the search query then using successive query refinements or alternatives. These query changes were made depending on the relevance of search results, i.e., relevance according to our intuitions about the query meaning and guided, where necessary, by the author's Phase One judgements. Our manual searches were not strictly manual in the same sense as the Cranfield 2 searches: we did use an automated search tool rather than search through papers by hand. We use the term 'manual' to indicate the significant human involvement in the searches. The manual searches took around 80 hours; they were a costly investment but we felt that they would be worthwhile, in order to find more relevant documents for the second round of judgements. We made liberal judgements, leaving the definitive judgements to the query author.

We then conducted some automatic searches. We ran the queries through three 'standard' IR models, implemented in Lemur[6]:

1. Okapi BM25 with relevance feedback (probabilistic model)
2. KL-divergence language modelling with relevance feedback (language modelling based model)
3. Cosine similarity (vector space model)

The intention behind using standard models from across the range of different retrieval model classes was to improve the reusability of the test collection. We did not include the

[6] http://www.lemurproject.org/

output from any of our citation methods; with so few models contributing to the pool, this could arguably create a bias in the test collection towards the methods we wish to evaluate. We were also constrained by time, since we wanted to minimise the delay between the authors making their first and second sets of judgements; we wanted their judgements to be as consistent as possible, in keeping with the laboratory model of relevance. Timing considerations were similarly taken into account when setting the parameters for the three models; we had neither the time nor enough relevance judgements from Phase One to optimise the models' parameters. We instead used what seemed to be reasonable parameter settings, based on a cursory review of the related literature and Lemur documentation. The parameter values we used are listed in Appendix B.5.

We pooled the manual and automatic search results, including all manual search results and adding one from each of the automatic rankings (removing duplicates) until 15 documents were in the list. If there were 15 or more manual search results, only these were included, as the manual search results were felt to be more trustworthy, having already been judged by a human as likely to be relevant. Our pool is very shallow compared to TREC-style pools; we rely on volunteer judges and therefore needed to keep the effort asked of each judge to a minimum.

The list of potentially relevant documents was then randomised and incorporated into personalised materials and sent to the query author with an invitation to judge them. We tried to make the task as easy as possible for the authors, to increase the likelihood that they would participate. The materials included instructions and a response form in both plain text and PDF, including the URL for a web page with identificatory details for the papers (i.e., title and authors) and links to the PDF versions of the papers, in order to aid the relevance decision. Again, sample materials are given in Appendix B.2.

We asked for binary relevance judgements in this second round, for the reasons discussed earlier and, also, in the hope that this would make the task easier for the authors and encourage a higher response rate. The instructions also asked authors whose research questions had been reformulated to approve our reformulations, i.e., to confirm that the reformulated query adequately represented their intended research question, and otherwise to give a more appropriate reformulation for resubmission to the pooling process.

3.3.4 Test collection statistics and analysis

In Phase One, out of around 315 invitations sent to conference authors, 89 resulted in research questions with relevance judgements being returned; 258 queries in total. Example questions are:

- *Does anaphora resolution improve summarization (based on latent semantic analysis) performance?*
- *Can knowledge-lean methods be used to discourse chunk a sentence?*

Of the 258 queries, 20 were from authors whose co-authors had also returned data. We treat queries from co-authors on the same paper as duplicates and use only the first author's. We discarded queries with no relevant Anthology-internal references but kept those whose only relevant references were intended for the Anthology but not yet included in the archive[7]. Queries with only one judged relevant paper in total, whether in the Anthology or not, were deemed too

[7]HLT-NAACL-2004 papers, for instance, were listed as 'in process' on the web site but were added later so could be included in our experiments.

Statistic	ACL Anthology			Other test collections			
	All Phase One	T_1	T_{1+2}	Cranfield 2	INEX 2005	TREC 8 Ad hoc	TREC Robust
# queries	196	82	82	221	63	150	50
Mean # judgements per query (Rel)	4.5	4.8	11.4	7.0	57	94	131.2
Mean # judgements per query (Irrel)	3.3	3.4	12.3	4.1	441	1642	624.74
# documents	9800	9800	9800	1400	17,000	528,000	1033,000
Mean # rel judgements per 1000 docs	0.46	0.49	1.16	5.00	3.35	0.18	0.13

Table 3.2: Test collection comparison.

specific and also discarded. In total, 61 queries were discarded due to these criteria, leaving 196 unique queries with at least one relevant Anthology reference and an average of 4.5 relevant Anthology references each. These are the queries in the first column of Table 3.2.

74 invitations were sent in Phase Two, totalling 183 queries. This is fewer than the 196 queries remaining after Phase One since 13 of the initially discarded queries were found too late to have (judged relevant) Anthology references, after Phase Two had been executed. These queries are included in the All Phase One set in Table 3.2. Similarly, a small number of discarded queries were mistakenly included in Phase Two. These queries are likewise included in All Phase One but not in the later sets: T_{1+2} is the complete test collection, i.e., the set of queries for which we have both Phase One and Two judgements and all those judgements. T_1 represents the T_{1+2} collection prior to Phase Two, i.e., the same queries but with only Phase One judgements.

44 Phase Two response forms were returned, giving judgements for 82 queries in total[8]. Appendix B.4 gives the final list of queries, which are identified by the ACL Anthology ID of the source paper, combined with the author's surname and a sequence number. 22 of these had been reformulated and all were approved by the author except two. In both cases, the author submitted an alternative reformulation for pooling and a new list (including the previous manual search results) was sent back for judgement. Both authors judged the (non-duplicate) documents in the new list.

Table 3.2 also compares our test collection to some other test collections. After Phase Two, the average number of judged relevant documents per query is 11.4; higher than for Cranfield 2, which had an average of 7.0 (Cleverdon et al. 1966). It is still low in comparison to, e.g., the TREC Ad hoc track, with an average of 94 judged relevant documents per query (Voorhees & Harman 1999).

However, the scientific aspect of the collection makes it very different in nature from TREC, with its news articles and related queries. Intuitively, because most scientific queries are very specialist, we do not expect a large number of relevant documents per query. A more appropriate modern comparison might be with TREC Robust (Voorhees 2005), whose queries are selected precisely for being 'hard', i.e., having few relevant documents. Furthermore, the document collection is also small in comparison to TREC and this possibly influences the absolute number of relevant documents per query. We have 1.16 judged relevant documents per thousand documents, compared with 0.18 for TREC 8 Ad hoc and 0.13 for TREC Robust[9]. Cranfield 2 has 5.00 judged relevant documents per thousand documents but these judgements are com-

[8] In fact, judgements were returned for 83 queries, including one discarded query with no relevant Anthology Phase One judgements, mistakenly processed in Phase Two.

[9] Counted from `http://trec.nist.gov/data/t14_robust.html`.

plete, made by searching the entire collection of 1400 documents, an infeasible task for modern collections, including our own.

Perhaps the closest modern comparison is with the INEX 2005 test collection, with its 17,000 IEEE articles; it has 3.35 judged relevant articles per thousand articles. However, the INEX judgements are made on a much deeper pool than we could realistically have used and it is probable that their judgements are closer to complete. Moreover, these document-level judgements are obtained from the original element-level judgements by simply treating any article containing a judged relevant element as relevant: they are not bona fide (document-level) judgements.

3.4 Chapter summary

The test collection paradigm raises some theoretical and practical issues for the IR experimenter; particularly for those intending to build a test collection. At the time when this work began, no available test collection was suitable for our experimental evaluation; we need a test collection with the full text of many cited and citing papers.

We have described how we created a new test collection around the ACL Anthology. The popular TREC-style methodology is too expensive for us; instead, we updated and adapted the Cranfield 2 methodology to our needs. Our queries are the research questions behind papers in the Anthology, as formulated by the paper authors themselves. Our relevance judgements are made by the query authors and include judgements on their paper's references and other papers in the ACL Anthology; we used pooling to find potentially relevant, non-cited papers in the ACL Anthology. The resultant test collection is small compared to other modern collections but we believe it to be a realistic and appropriate collection for our purposes.

Grade	**Description** and examples
4	**The reference is crucially relevant to the problem. Knowledge of the contents of the referred work will be fundamental to the reader's understanding of your paper. Often, such relevant references are afforded a substantial amount of text in a paper e.g., a thorough summary.** • In the case of subproblems, the reference may provide a complete solution (e.g., a reference explaining an important tool used or method adopted for the research). • In the case of the main problem, the reference may provide a complete solution (e.g., an existing, alternative solution to the problem that your work directly contrasts or proves incorrect). • In either case, the reference may provide a partial solution that your work builds upon (e.g., previous work of your own or others that your current work extends or improves).
3	**The reference is relevant to the problem. It may be helpful for the reader to know the contents of the referred work, but not crucial. The reference could not have been substituted or dropped without making significant additions to the text. A few sentences may be associated with the reference.** • The reference may be the standard reference given for a particular tool or method used, of which an understanding is not necessarily required to follow your paper. • The referred work may give an alternative approach to the problem that is not being directly compared in the current work. • The referred work may give an approach to a similar or related problem.
2	**The reference is somewhat (perhaps indirectly) relevant to the problem. Following up the reference probably would not improve the reader's understanding of your paper. Alternative references may have been equally appropriate (e.g., the reference was chosen as a representative example from a number of similar references or included in a list of similar references). Or the reference could have been dropped without damaging the informativeness of your paper. Minimal text will be associated with the reference.** • The reference may be included to give some historical background to the problem. • The reference may be included to acknowledge a (non-critical) contribution.
1	**The reference is irrelevant to this particular problem.** • E.g., a reference about an implementation strategy may be irrelevant to a subproblem about evaluation strategy.

Table 3.3: Graded relevance scale.

Chapter 4

Document collection

In our experiments, we propagate text from citing papers to cited papers to be used as additional index terms. Therefore, references from documents in the test collection to other test collection documents will be most useful. We call these *internal* references. It is practically impossible to find or create a collection of documents with only internal references but the higher the proportion of these, the more useful the test collection will be for citation experiments. We discussed in Chapter 3 why real, naturally occurring document collections are preferrable in retrieval experimentation; to build our test collection, we therefore looked for an existing document collection, from a relatively *self-contained* scientific field. When choosing a field to study, we looked for one that is practicable for us to compile the document collection – freely available machine-readable documents; as few as possible document styles – while still ensuring good coverage of research topics in the entire field. Had we chosen the medical field or bioinformatics, for example, the prolific number of journals would have been a problem for the practical document preparation.

Computational linguistics (CL) is a small, homogeneous research field and one that we intuitively recognise to be fairly self-contained. The ACL (Association for Computational Linguistics) Anthology is a freely available digital archive of CL research papers[1]; it contains the most prominent publications since the beginning of the field in the early 1960s, consisting of only one journal, six conferences and a few other, less important publications, such as discontinued conferences and a large series of workshops. Table 4.1 lists these publications and their official identifiers. The archive totals more than 10,000 papers[2]. In the ACL Anthology, we expect a high proportion of internal references within a relatively compact document collection.

We empirically measured the proportion of Anthology-internal references, using a sample of five papers from each of five of the Anthology's main publications. We found a proportion of internal references to all references of 33.00% (the *in-factor*). Table 4.2 shows the in-factor within each of the five publications. We wanted to compare this number to a situation in another, larger field (namely, genetics) but no straightforward comparison was possible[3], as there are very many genetics journals and quality of journals probably plays a larger role in a bigger field. We tried to simulate a collection that is similar to the seven main publications in the Anthology, by considering a range of fixed groups of genetics journals. We used the ISI Journal

[1] `http://www.aclweb.org/anthology/`

[2] This is our estimate, after substracting non-papers such as letters to the editor, tables of contents etc. At the time of writing, the Anthology web site reports that it contains 12,500 'papers'. The Anthology is growing by around 500 papers per year.

[3] At the time our work began, the full-text TREC Genomics collection was not yet available.

ID	ACL Anthology publication
P	Proceedings of the Annual Meeting of the ACL (ACL)
N	Proceedings of the North American Chapter of the ACL (NAACL)
C	Proceedings of the International Conference on Computational Linguistics (COLING)
W	Proceedings of various ACL workshops
H	Proceedings of the Human Language Technology Conference (HLT)
J	Computational Linguistics Journal
E	Proceedings of the European Chapter of the ACL (EACL)
A	Proceedings of the Applied Natural Language Processing Conference (ANLP)
I	Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)
T	Proceedings of the Theoretical Issues in Natural Language Processing Conference (TINLAP)
X	Proceedings of the Tipster Text Program
M	Proceedings of the Message Understanding Conference (MUC)

Table 4.1: ACL Anthology publications and their identifiers.

ACL Anthology publication	% Internal references
Computational Linguistics Journal	18.41%
ACL Proceedings	33.54%
COLING Proceedings	42.70%
HLT Proceedings	37.26%
ANLP Proceedings	33.10%
(Mean)	(33.00%)

Table 4.2: Proportion of internal references in ACL Anthology papers.

Citation Reports's Genetics & Heredity subject category as our definition of the field and took samples of five papers from various subsets of these 120 journals[4]. We varied the following factors in our subsets:

1. Range of journal impact factors[5] (Mixed vs Top)
2. Number of journals (5 vs 10)
3. Definition of internality (Local vs Global)

In the Mixed subsets, journals ranging from high to low impact factor were sampled whereas, in the Top subsets, the top journals as ranked by impact factor were taken. The Top subsets were intended to simulate the fact that the Anthology contains most of the prominent publications in the field: impact factor is an indicator of the relative 'importance' of a publication. Likewise, the Mixed subsets simulate how the Anthology covers most of the *total* publications in the field, i.e., from across the range of importance. Local is the stricter of our two definitions of reference internality, meaning a reference to a journal within the same subset. Global internality is more

[4] `http://scientific.thomson.com/products/jcr/`

[5] Journal impact factor is a measure of the frequency with which its average paper is cited and is a measure of the relative importance of journals within a field (Garfield 1972).

Journal subset	% Internal references	
	Local	Global
Mixed 5	4.35%	13.60%
Mixed 10	5.13%	16.81%
Top 5	7.00%	20.81%
Top 10	13.13%	22.70%

Table 4.3: Proportion of internal references in genetics papers.

liberal, meaning a reference to *any* journal in the ISI Genetics & Heredity category. Table 4.3 gives the in-factors within each of these subsets, using both the local and global definitions of internality.

The highest in-factor measured was 22.70%, from the Top 10 journal subset, using global internality; this is lower than the 33.00% measured for the Anthology and could only be achieved by an impracticably large collection: there are 120 journals in the ISI list. This supports our hypothesis that the Anthology is reasonably self-contained, at least in comparison with other possible collections. The choice of computational linguistics has the additional benefit that we are familiar with the subject matter and can better analyse and interpret experimental results (i.e., retrieval results) using our knowledge of the field; better than we would be able to in, e.g., the biomedical domain.

Hence, we centred our test collection around the ACL Anthology. Our document collection is a $\sim$9800 document snapshot of the archive; roughly, all documents published in 2005 or earlier, with non-papers (e.g., letters to the editor) removed. Anthology document identifiers are of the form 'J00-1002', where the 'J' indicates the document is from the Computational Linguistics journal, the following '00' is the last two digits in the document's publication year (2000) and the four digit number following the hyphen is a unique identifier within the J00 documents. Figure 4.1 lists the identifiers for the publications archived in the ACL Anthology.

Chapter 5

Document processing and citation database

The ACL Anthology documents are archived in Portable Document Format (PDF), a format designed to visually render printable documents, not to preserve editable text. In order to conduct our experiments with citation information, using the text from around the citations, the PDF documents must be converted to a fully textual format and processed to identify and access the pertinent information. A pipeline of processing stages was developed in the framework of a wider project, illustrated in Figure 5.1; the final two processing stages, in dotted outline, are those developed for this thesis work.

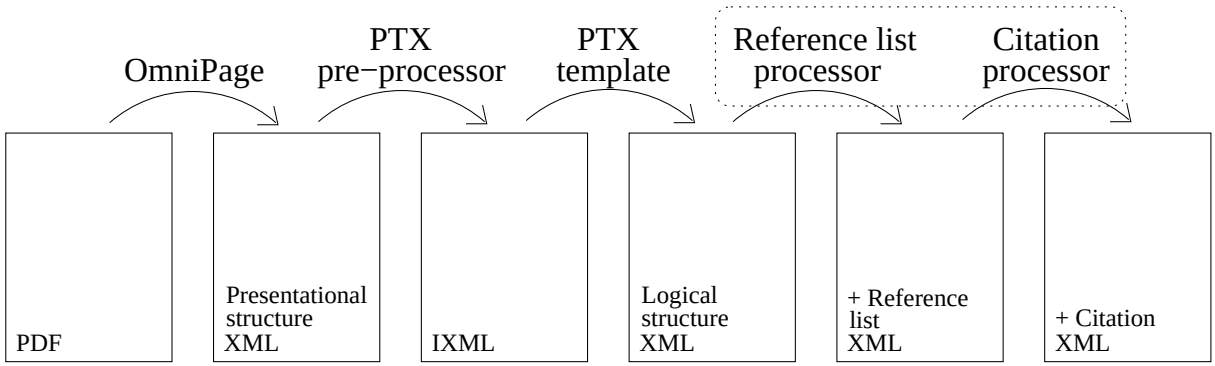


Figure 5.1: Document processing pipeline.

Firstly, OmniPage Pro 14[1], a commercial PDF processing software package, scans the PDFs and produces an XML encoding of character-level page layout information. AI algorithms for heuristically extracting character information (similar to OCR) are necessary since many of the PDFs were created from scanned hard copies and others do not contain character information in an accessible format. The OmniPage output describes a paper as text blocks with typesetting information such as font and positional information. Next, a software package called PTX (Lewin, Hollingsworth & Tidhar 2005) first filters and summarises the OmniPage output into Intermediate XML (IXML), as well as correcting certain characteristic errors from that stage. Then, a publication-specific template converts the IXML to a logical XML-based document structure (Teufel & Elhadad 2002), by exploiting low-level, presentational, style information such as font

[1] http://www.scansoft.com/omnipage/

1. Trevor Strohman, Donald Metzler, Howard Turtle, and W. Bruce Croft. 2005. Indri: a language-model based search engine for complex queries. Technical report, University of Massachusetts.

2. Simmons, R. and Slocum, J. 1972 Generating English discourse from semantic networks. *Communications of the ACM* **15** (10) October, 891-905.

3. [11] Simmons, R., and J. Slocum, "Generating English Discourse from Semantic Networks," *Comm. ACM 15, 10* (October 1972), 891-905.

Figure 5.2: Reference examples, taken from (1) N06-1050, (2) J82-2003 and (3) J81-1002.

size and positioning of text blocks. As part of this processing stage, PTX tries to identify where the reference list starts and ends, and where the individual references in the list start and end; the XML document output by PTX thus includes a basic annotation of the reference list.

The subsequent processing stages were developed by the present author for the purposes of this research and are described in more detail in the following sections. They incrementally add more detailed information to the logical representation. Firstly, the paper's reference list is annotated in more detail, marking up author names, titles and years of publication in the individual references. Finally, a citation processor identifies and marks up citations in the document body and their constituent parts, e.g., author names and years, before associating each citation with the corresponding item in the reference list. Extracting this citation and reference information is a non-trivial task, for which high precision methods have been developed independently (Powley & Dale 2007). Once the documents are processed, we extract the citation information required for our retrieval experiments to a database.

5.1 Terminology

At this point, a note about terminology is appropriate. A citation is a directed relationship between two documents: when document A cites document B, there is a *reference* to B in A and B has received a *citation* from A. Strictly speaking, a document's citations means the incoming acknowledgements it receives from other documents; the outgoing acknowledgements that a document gives to others are its references. In this thesis, we are concerned with both citing and cited documents, with references and citations; in general, we try to maintain this terminological distinction. However, in this chapter, we discuss a document's references at the typographical level and, in particular, we distinguish between the full textual reference listed in the bibliography (or reference list) at the end of the document and the textual markers in the document's running text that show at which specific points a reference is being cited. When this is the case, we will reserve *reference* for the former and misappropriate the term *citation* for instances of the latter.

5.2 Reference list processing

Referencing is a standardised procedure, where certain pre-specified information is given about a cited work. Nevertheless, there is a great deal of variation among references. Firstly, there are references to different types of document, for which slightly different information must be given. Secondly, there are many different reference styles, between which the ways that information is presented can differ greatly. Consider the example references in Figure 5.2, reproduced exactly as they appear in the ACL Anthology papers from which they are taken. The first is a reference for a technical report, consisting of authors, publication year, title and institution, whereas the second is for a journal article, also with author, publication year and title but then journal name, volume, number and page numbers. The third example is another reference to the same journal article but in a different style, where the reference is numbered, the same information is presented in a different order, the journal name is abbreviated and the punctuation, capitalisation, italicisation and emboldening conventions are different. In all three references, the format of the sequence of author names is different. Thus, the task of automatically extracting bibliographic information from reference lists is non-trivial. In the input to our reference list processor, the reference list is already segmented into individual references. Our method uses heuristics and a library of regular expressions developed by a thorough study of the reference styles in ACL Anthology papers; even in such a limited number of publications, the range of styles is considerable. Rather than attempt an exhaustive annotation of each reference, we search only for those pieces of information that a) we need to extract for later processing stages and/or b) are useful for identifying those required pieces of information. Namely, we tag the reference title, the author name(s) (and, in particular, the surname(s)) and the date of publication.

For each reference, the processor begins by searching for strings that look like a publication year. Next, the string preceding the publication year is searched for a list of author names. Thirdly and finally, the reference title is tagged in the string immediately following the publication year.

5.3 Citation processing

First appearances suggest that the textual format of citations is more restricted than that of references: there are three broad categories of citation style, which we will call *numeric*, *abbreviated* and *nominal*. In the numeric style, each item in a reference list is numbered and citations in the running text are simply bracketed numbers that correspond with the list. Similarly, in the abbreviated style, each reference is given an identifier formulated from, e.g, the author names and publication year, which is used as the citation, e.g., [Cah92]. These two styles are rarely seen in ACL Anthology papers, however, where the nominal style is prevalent. Here, citations roughly consist of bracketed names and dates. Yet there is still much scope for complication within this style, as illustrated by the examples in Figure 5.3, again, taken from ACL Anthology papers.

Citations can be *parenthetic* or *syntactic*, depending on whether the citation is a necessary grammatical constituent in the sentence in which it occurs, i.e., whether it syntactically functions as a noun phrase. This is sometimes distinguished textually by bracketing the entire parenthetic citation (as in example 1 of Figure 5.3) or leaving the author names outside the bracketed part for syntactic citations (as in 2). Lists of author names can be conjoined with an ampersand (as in example 3), instead of *and*, or lists may be abbreviated using *et al* (example 4), though this abbreviation may only occur in the second and later citations to that reference in

1. (Bikel, 2004) [taken from W05-1528]
2. Grosz, Joshi, and Weinstein (1995) [J99-4007]
3. (articulated in Kintsch & van Dijk [1978]) [J99-4006]
4. (Marcus et al., 1993) [P03-2036]
5. Cohen's model (1981) [J95-3003]
6. (for example, McCord, 1990; Hobbs and Bear, 1990) [J94-4005]
7. Pollard and Sag (1994, p. 360, fn. 20) [J97-4003]
8. (Grosz, Joshi, and Weinstein, 1995 henceforth GJW)...
   GJW (1995, p. 215, footnote 16) [J97-3006]
9. (Charniak, 1997; Collins, 1997, 2000; Eisner, 1996) [P04-1058]
10. Prince's (1981; 1992) [P98-2204]
11. (Grosz 1977a; Grosz 1981) [J99-4006]
12. (Kameyama 1986; Brennan, Friedman, and Pollard 1987; Di Eugenio 1990, 1996; Walker, Iida, and Cote 1994; Strube and Hahn 1996, inter alia; see also citations within GJW, forthcoming papers in Walker, Joshi, and Prince in press, and psycholinguistic studies described in Hudson-D'Zmura 1989, Gordon, Grosz, and Gilliom 1993, and Brennan 1995) [J97-3006]

Figure 5.3: Citation examples.

the text. Example 3 also illustates how the type of the brackets can differ, especially when citations interact with other parentheticals in the text. Syntactic citations can occur as part of possessive noun phrases, e.g., examples 5 and 10. Additional strings can also appear both before the names (example 6) and after the dates (example 7). References which are cited frequently in a given paper may be cited in full once, introducing an abbreviation for the citation to be used from then on, as in example 8. Citations to multiple publications can appear as sequences within the same brackets (example 9). When publications in the same citation group are by the same author(s), the author list may be presented only once, while the publication dates are listed in sequence (example 10), or each citation may be presented in full (example 11). This example also shows how citations to different publications which share the the same author list and publication date may be distinguished using additional characters concatenated to the publication dates. Authors sometimes typeset/format their citations manually rather than using a bibliographic software tool, which can result in errorful and/or non-standard citations. Finally, any combination of these features may occur together, to form some hugely complex citations (example 12).

In order to be able to automatically recognise instances of such a complex phenomenon in our documents, we first conducted a detailed study of the citation formats in ACL Anthology documents. We next developed a comprehensive grammar of regular expressions for textual

citations. Figure 5.4 gives some example regular expressions from our citation grammar, reproduced in full in Appendix C. Our citation processor begins by extracting all author surnames from the annotated reference list to a lexicon. This lexicon is used to search for and annotate instances of those surnames in the parts of the document that are of interest, e.g., title, abstract, paragraphs in the body of the text, footnotes etc. Finally, beginning from the annotated surnames, the citation grammar is used to search for text that looks like constituent parts of a citation, annotate them and combine them into larger constituents and, eventually, a complete citation. In a first post-processing stage, the textual citation annotations are converted to a logical annotation, i.e., citation sequences are separated into individual citations, each with their own annotations, including duplicating multi-paper citations with the same authors, like example 10. Finally, the citations are compared with the reference list to find their corresponding reference, and attributed with identifiers accordingly.

```
our $SYNTACTIC =
  '()(' . $AUTHOR . '(\s)*(' . $COMMA . ')?(\s)*(' . $DATEHENCEFORTHPC . ')((\s)*' . $POSTSTRING . ')?)()';

our $PARENTHETIC =
  '()(' . $LBR . '(\s)*(' . $PRESTRING . ')?(\s)*(  . $AUTHORSIMPLEDATECOMMA . ')+(\s)*(' . $POSTSTRING .
  ')?(\s)*' . $RBR . ')()';

our $PRESTRING =
  '()(' . $PRESTRINGWORD . '(((\s*' . $COMMA. ')?\s*' .$PRESTRINGWORD . ')+)?(\s*' . $COMMA. ')?)()';

our $PRESTRINGWORD =
  '(\b)((see\b)|(also\b)|(e(\.)?(\s)*g(\.)?)|(in\b)|(for example\b)|(such as\b)|([cC](\.)?f(\.)?))()';

our $AUTHOR =
  '()(' . $NAMEETAL . '|' . $NAMES . ')()';

our $NAMEETAL =
  '()(' . $NAME . '(\s)*?(' . $COMMA . ')?(\s)*?' . $ETAL . '(\s)*?(' . $FULLSTOP . ')?)((\s)*?(' . $GENITIVE . ')?)';

our $ETAL =
  '(\s+)([Ee][tT](?:\.)?(?:\s)*?[Aa][lL](?:\.)?)(\s*)';

our $NAME =
  '()((' . $PRENAME . '(\s*))?' . $SURNAME . ')()';

our $PRENAME =
  '(?:\b)(?:[Ll]a|[Dd][iue]|[dD]ella|[Dd]e\s+[Ll]a|[Vv]an\s*?(?:[td]e[rn])?)(?:)';

our $SURNAME =
  '()(<SURNAME>[^<]*?</SURNAME>)()';
```

Figure 5.4: Example regular expressions from citation grammar.

5.4 Sentence segmentation

As a final document processing step, we use the tokeniser from a statistical natural language parser (Briscoe & Carroll 2002) to segment the text into sentences. Sentence boundary detection is an important pre-processing task for many natural language processing tasks, such as machine translation (Walker, Clements, Darwin & Amtrup 2001). The task is non-trival: basic cues like sentence-terminal punctuation (e.g., '.', '?', '!', and ':') and start-of-sentence capitalisation are complicated by token-internal punctuation (in numbers, times, abbreviations etc.) and language-specific capitalisation conventions. The sentence boundary detector, in our case, is helped somewhat by PTX's segmentation of the paper into paragraphs, using visual

page layout information. However, this approach sometimes incorrectly splits sentences that occur across page boundaries. Nevertheless, the input to the boundary detector is of fairly good quality. The nature of the documents – well written, highly edited text – makes the task slightly easier than, for example, on automatic speech recognition output. The task is further simplified by a pre-processing step which enables the XML-tagged citations to pass through the parser software as grammatical tokens[2]; this circumvents potential problems from punctuation within the citations. Overall, the output from this processing stage is satisfactory: the majority of segments inspected are complete, grammatical sentences and the remainder are mostly corrupted by PTX errors, rather than incorrect sentence boundary detection.

5.5 Citation database

The final major task in preparation for our retrieval experiments is to build a database of the necessary citation information. Figure 5.5 is a schematic overview of what a record in the database contains and how it is created. We use an open source XML database with XQuery-based access, called Oracle Berkeley DB XML[3], and initially populate our database with bibliographic information about the ACL Anthology papers extracted from the Anthology web site's HTML index pages[4], as well as their unique ACL Anthology identifiers. We add some further information at this stage, discussed shortly.

We next identify which references in our documents are to other documents in the ACL Anthology, by searching for strings in the references that correspond to the name of an Anthology publication. We compare against a library of publication names which we manually constructed from an inspection of the ACL Anthology web pages and our knowledge of common abbreviations for the main publications, given in Appendix C. So, since the reference in our illustrated example contains the string *Computational Linguistics*, it is identified as a reference to a CL journal paper. For each of those references, we then find the citations in the running text that are associated with the reference (i.e., those with a matching identifier attribute) and extract words from around those citations to our database. In fact, we extract a variety of citation contexts, i.e., words from a range of extents around the citation, and add each of these usually as a separate field to the database record. As an exception to this, we create a single field consisting of the sentence containing the citation plus four sentences on either side of the citation; from this one field, we create a range of fixed window contexts. Our default citation context is the sentence that contains the citation. In this case, for example, we extract the entire sentence to the database except for the textual citation itself, its XML tags and any other citations, all of which we remove. The full range of contexts will be discussed in Chapter 6.

In order to add these citation contexts to our database, we must determine which database record (c.f. ACL Anthology paper) the citation refers to. In principle, this is straightforward: our database contains the bibliographic information for the paper (its title and author names), which is generally enough to distinguish the paper from any others. However, several factors mean that the task is more complicated than could be solved by simple string matching on these fields. Firstly, there are occasional errors in the Anthology index files. Secondly, there are sometimes errors in the references from the earlier document processing stages, e.g., OmniPage character recognition errors, incorrect PTX segmentation of the reference list and/or incorrect annotation of the reference by the reference list processor. Even with perfectly processed references,

[2]Many thanks to Don Tennant for developing this perl script.

[3]`http://www.oracle.com/database/berkeley-db/index.html`

[4]E.g., `http://aclweb.org/anthology-new/J/J93/index.html`

the reference information need not match that of the (usually) correct index file information, which we take as normative. References can contain typographical errors by the paper authors and, for some papers, bibliographic information may be presented inconsistently. For instance, long titles maybe be abbreviated in some references, lists of authors may be truncated or author names may appear slightly differently, e.g., due to marital name changes.

Thus, we require a more robust solution than naive string matching. Our solution is based on work done on duplicate detection in databases. Duplicate records can occur, for example, when multiple databases are merged, wasting space and potentially causing more serious problems. In large databases, the cost of comparing every pair of records for duplicity is far too high and, instead, techniques have been developed to bring putative duplicates together. Then, more expensive comparisons can be carried out on these much smaller sets of similar records. The task of detecting duplicates in bibliographic databases, in particular, is very much like our own problem: given a bibliographic record, they are trying to detect any other records that contain matching bibliographic information; given a reference, we are trying to 'detect' a single bibliographic record that matches the bibliographic information from that reference. Ridley (1992) notes the problems that slight title differences etc. cause for simple techniques to bring duplicates together.

Our method is based on Ridley's (1992) 'expert system' for duplicate detection in bibliographic databases. Their system makes use of the Universal Standard Bibliographic Code (USBC), a fixed length code comprised of elements representing various bibliographic information (Ayres, Nielsen, Ridley & Torsun 1996). This code was invented as a universal standard book number, to uniquely identify books; 'universal' because it is created by a logical process on their bibliographic cataloque entries that would generate the same control number in any computer environment. The USBC was developed to obtain maximum discriminatory power from as short a code as possible. Ridley's (1992) system firstly creates clusters of records with very similar bibliographic information, i.e., the same USBC. In their database of nearly 150,000 records, most of these clusters contained only two or three records. Finally, an 'expert' set of manually developed tests are conducted on the clustered records, to determine whether they are true duplicates, i.e., whether they represent the same bibliographic entity.

Our method proceeds similarly. For each reference, we generate a reduced version of the USBC, created solely from the title of the reference. This is created by concatenating the seven least frequent alphanumeric characters in the reference title, in ascending order of frequency, after converting to uppercase; for our illustrated example, the code FPLCHOR is created from the title string *The Mathematics of Statistical Machine Translation: Parameter Estimation*. Because this code is so much shorter than the full title and consists only of the rarest and, hopefully, most distinguishing characters in the title, discounting whitespace and punctuation, the title USBC will be more robust to variations in the exact form of the title. We use only the title element of the USBC because, in a moderately sized database like ours, the probability of multiple papers having similar enough titles to generate the same title code is small enough that the distinguishing power of this element alone is great enough for our purposes; including, e.g., author and publication date codes would only increase the scope for introducing errors into the code, resulting in failure to match references with the correct database record.

For the majority of our references, the generated title USBC matches that of a single database record; an inspection of these matches showed that this is a clear, reliable indicator that the correct record has been found. In the unusual case where there is more than one match, we attempt an exact string match on the full reference title; if there is a single match, we accept

this as the correct record; if there is more than one, we compare the publication date and which publication the cited paper comes from to the same information from each of the title-matching records and only accept the record which matches both. For some references, the title USBC does not generate any match in the database. In these cases, checking for an exact string match on the title is futile: an unmatched USBC means that there must be some discrepancy between the reference title and that of the correct database record. A more complex match could be attempted, e.g., some fuzzy match on the title or title USBC, and/or using other bibliographic information. However, these cases are sufficiently few and we obtain sufficiently many successful matches from title USBCs alone, for the purposes of this work, that we leave these cases unresolved.

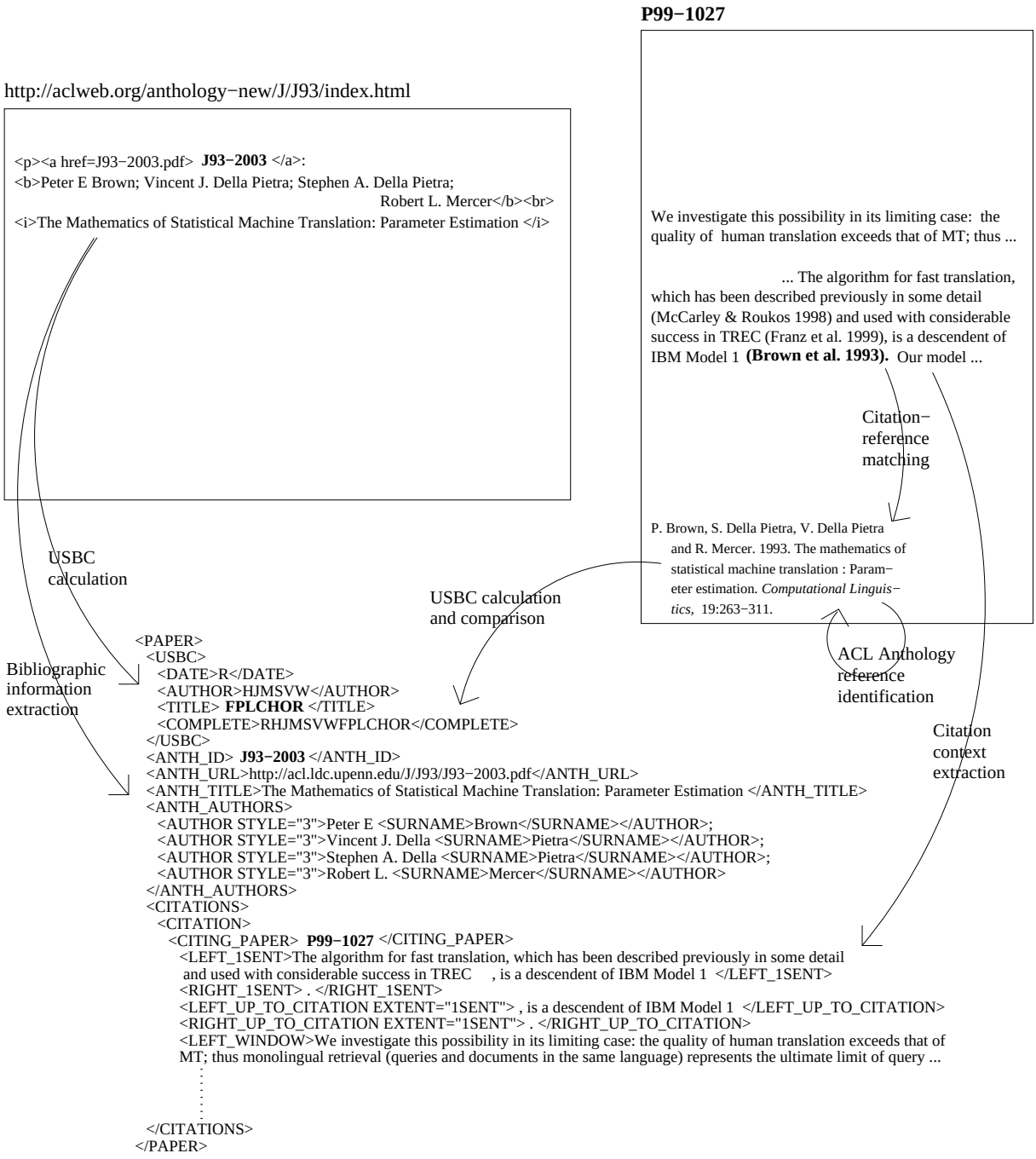


Figure 5.5: Construction of citation database record.

Database statistics

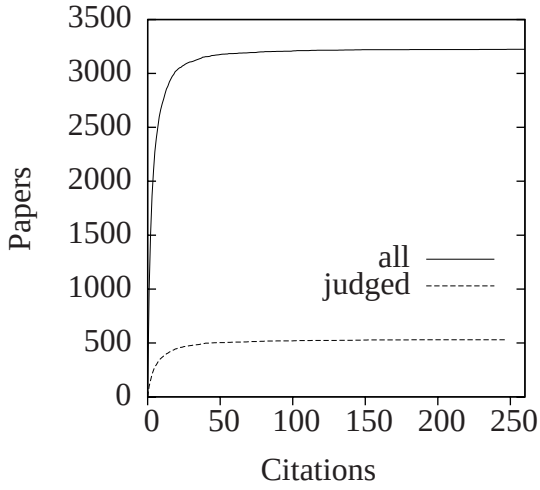


Figure 5.6: Cumulative total of papers with citations by number of citations per paper.

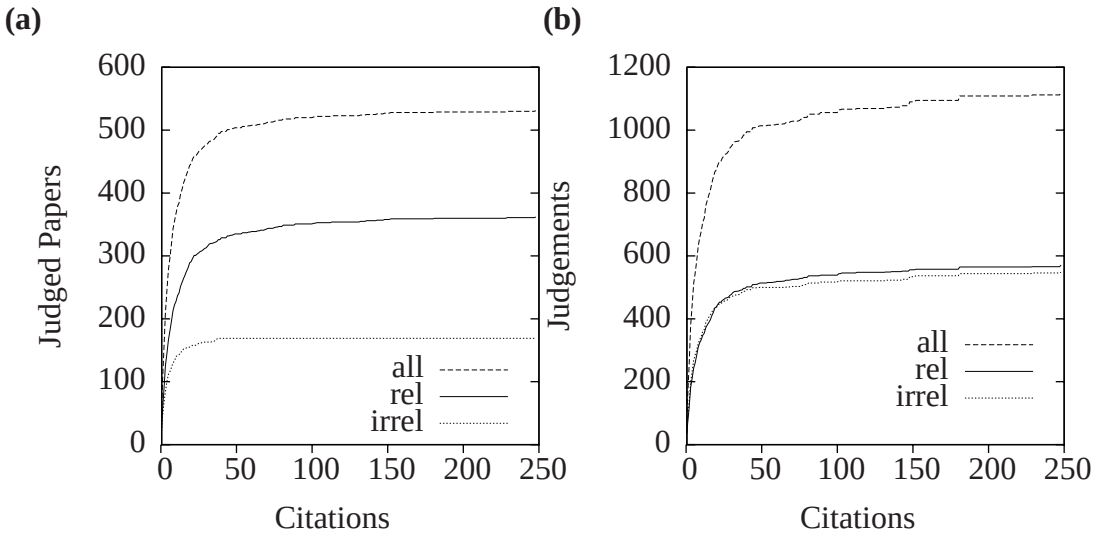


Figure 5.7: Cumulative totals of papers with citations by number of citations: **(a)** judged papers and **(b)** individual judgements.

Our database contains contexts from over 20,000 citations to over 3200 papers. Figure 5.6 shows how this total number of citations is distributed over the papers, as a cumulative total, according to the number of citations each paper has: the y-value of a given point on the line shows how many papers have x or fewer citations in the database. The curve tails off rapidly as the number of citations increases: almost one thousand papers have only a single citation in the database; a single paper has 248 citations, the greatest number in the database. The majority of papers in the database (over 7000) have no citations, in keeping with the idea that only a minority of influential papers are eventually cited. The lower, dotted curve in Figure 5.6 is the equivalent plot for only those papers for which we have relevance judgements in our

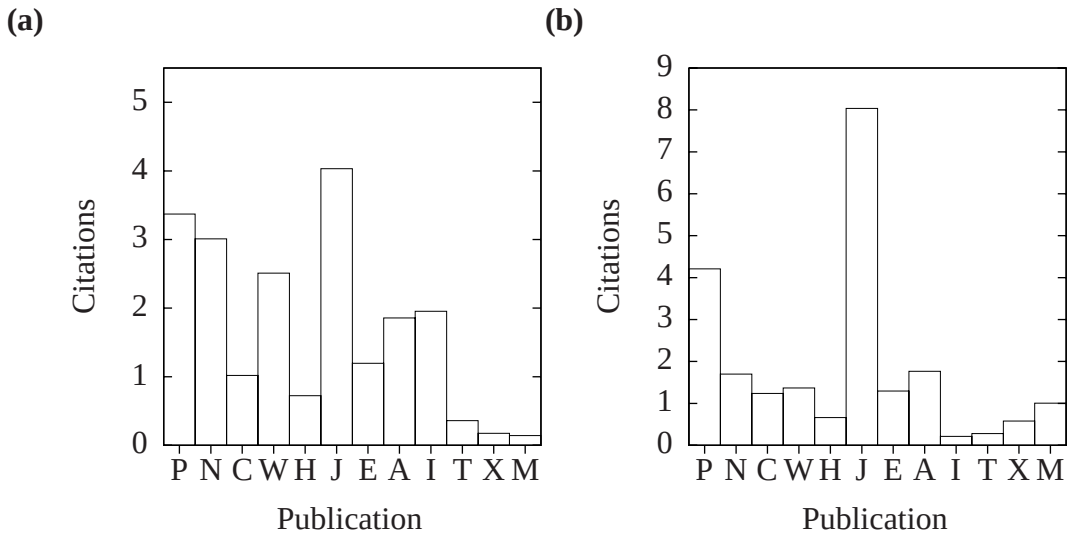


Figure 5.8: Distribution of citations by publication of (a) citing and (b) cited paper, normalised by the total number of papers from that publication.

test collection; it is citations to these papers which we anticipate to be most influential in our experiments. Naturally, there are far fewer citations to judged papers, since the majority of papers were not judged with respect to any query.

Figure 5.7(a) shows the same cumulative plot for judged papers in closer detail and also decomposes this curve into one for papers which have been judged relevant for some query and another for papers which are judged irrelevant for every query for which they were judged. Most judged papers are judged with respect to more than one query, because our query authors often submitted multiple related research questions as queries. We see that the 'judged relevant' curve is higher: most judged papers are judged to be relevant for at least one query.

Citations to judged papers will particularly influence the results of each one of the queries for which it is judged. To give a perhaps more accurate idea, then, of how much 'influential' citation data we have in our database, Figure 5.7(b) gives the equivalent plots for individual relevance judgements, rather than judged papers: we have citations for over 1000 judgements, including almost 600 'relevant' judgements. The difference between the 'judged relevant' and 'judged irrelevant' curves is much smaller here: most papers, though they are judged relevant for one query, are not judged relevant for all the queries for which they are judged.

In Figure 5.8, we look at how the citations are distributed with respect to publication of the citing and cited papers. The publication IDs are explained in Table 4.1, Chapter 4. The number of citations is normalised by the total number of papers in that publication, since this varies greatly between publications. We see that, according to our database, papers from the Computational Linguistics journal (J) cite other Anthology papers more often than any of the other publications do. This does not necessarily indicate a higher in-factor; it is probably a product of the greater average number of references (and citations) in journal papers. However, journal papers are also cited most often out of all the publications, suggesting that journal papers are typically the most important or influential, compared to conference and workshop papers.

Figures 5.9(a) and 5.9(b) show the distribution of the citations by year of the citing and cited paper, respectively. This illustrates how most of the citations within the Anthology occurs between the more recent papers. It is unsurprising that we find very few Anthology-internal

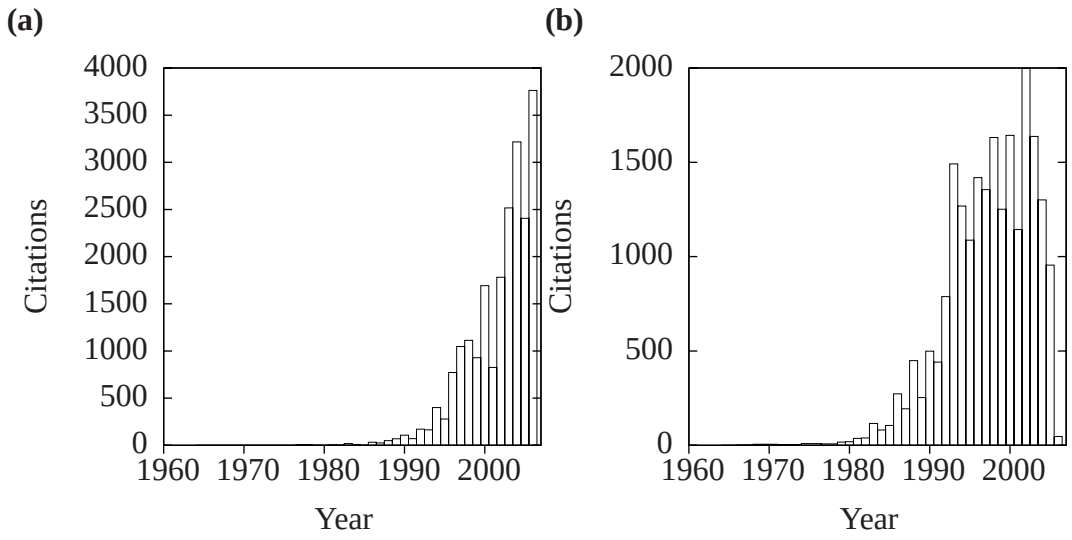


Figure 5.9: Distribution of citations by year of (a) citing and (b) cited paper.

citations *from* old papers: papers can only ever cite earlier work so, the older the paper is, the older its references will be and the more unlikely it is that those references will also be from the period covered by the Anthology. It follows that, the older a paper is, the greater the number of subsequent papers there are that *could* cite it; one might expect that papers from the earlier years of the Anthology would be the most heavily cited. However, Figure 5.9(b) shows the opposite: the vast majority of the citations are to the most recent papers. This is in keeping with observations from citation analysis that authors tend to cite the most recent work that is relevant to their own, rather than citing all historically influential works (see, e.g., Garfield 1997).

5.6 Evaluation and comparison with other work

We conducted an intensive evaluation on 10 CL journal papers, inspecting the eventual output of our processing and identifying in which of the successive stages along the way errors are made. Overall, our document processing performs well: we find and correctly match 388 out of 461 citations with their corresponding reference (84.2%). Errors mostly occur due to noise from the PDF to XML conversion prior to our processing, e.g., OmniPage character recognition errors and incorrect segmentation of the reference list by PTX. Other methods for automatically identifying reference information have been developed independently of our work. Powley & Dale (2007) extract the same information from Anthology PDF papers; Powley & Dale convert the PDFs to plain text using the open source tool PDFBox[5], then take a similar approach to ours, looking for instances of author surnames in the reference list to confirm the authenticity of candidate citations. They do not report a comparable evaluation for the overall task but, for instance, report precision of over 99% and recall of 91% on the citation-reference matching stage, in an evaluation on 60 Anthology papers; the equivalent numbers from our evaluation are 99% precision and 92% recall. The ParsCit tool (Councill, Giles & Kan 2008) uses machine learning to perform a more detailed annotation of reference lists from plain text input, e.g., identifying journal names, editors and locations, as well as the author names, dates and titles we annotate. In another study, of 30 Anthology papers this time, ParsCit achieved precision and

[5] `http://www.pdfbox.org/`

recall of 85% and 85% on author names, for example, compared to our 98% and 94%[6].

Out of the 461 citations, 192 are to Anthology papers; we end up with a citation context in our database for 106 of these (55.2%). These 192 citations correspond to 87 references to Anthology papers; we find citation contexts for 44 of these (50.6%). The task of identifying which references are to Anthology papers is a hard task in itself, even for humans. In the ACL Anthology Network project (Joseph & Radev 2007), student research assistants manually inspected the reference lists of the Anthology documents and created a list of Anthology-internal citation links; in the corresponding entries for our 10 journal papers, only 85 of the 87 Anthology references (97.7%) were identified. Our 50.6% is the result of several stages of automatic processing, all of which introduce some error. Again, the majority of the 43 'misses' (32 of them, 74.4%) are due to errors in the PDF to XML processing stages, with a further two caused by errors in our own document processing (4.7%). This leaves 10 references out of 87 for which we fail to extract citation contexts at the database stage (11.5%).

We also attempted a larger comparison of the citation links in our database with the ACL Anthology Network list, taking the latter as the gold standard. The comparison is not straightforward for two reasons. Firstly, some of the paper IDs in our database are not official Anthology IDs, since those papers were only assigned their official IDs after we had carried out our document processing and we used the temporary IDs they had when they were first distributed. Secondly, the ACL Anthology Network list and our database are not created from identical sets of documents: the Network list was created later than we took our snapshot of the archive, after it had continued to expand. If we compare against the entire Network list, we arrive at precision and recall values of 85.8% and 18.1% for our database links, i.e., out of 8174 database links (from 3979 citing papers), 7017 match an entry in the gold standard, of which there are 38,765 (from 8437 citing papers). If we only consider the gold standard links from the 3979 citing papers in our database, however, recall is 32.2%. Precision is probably higher than 85.8% since at least 757 of the 1157 database links that do not match a gold standard link are from papers with non-official IDs; if all 757 of these match a gold standard link, precision would be 95.1%.

[6]Thanks to Awais Athar, who conducted this evaluation as part of his MPhil project.

Chapter 6

Experiments

Having discussed the preparation of our experimental data, we now recapitulate what experiments they are intended for. We experiment with the combination of terms from citing and cited documents, a condition that is not commonly tested; previous work has generally tested one or the other. Our test collection, with the full text of a substantial number of citing and cited documents, allows broad experimentation with combinations of information from the citing and/or cited documents. In our experiments here, we take words from around citations in citing documents and add those to a base representation of the cited document, i.e., the entire cited document. In the following section, we discuss our experimental set-up; the methods and tools we use in creating our combined document representation, indexing our documents, running our queries against the index and evaluating the retrieval results. In Section 6.2, we discuss the evaluation measures we use. Then, Section 6.3 presents the results of some preliminary retrieval runs, with the intention of establishing the validity of our test collection as an evaluation tool. Here, we also introduce the notation we use for incorporating statistical significance information into system rankings. Sections 6.4 through 6.6 present our main experiments: first, the basic experiments comparing retrieval effectiveness with and without citation terms; next, experiments where citation terms are weighted higher relative to document terms; finally, experiments comparing a range of contexts from which citation terms might be taken.

6.1 Method and tools

We index our documents using Lemur, specifically Indri (Strohman, Metzler, Turtle & Croft 2005), its integrated language-model based component, using the SMART stoplist (Buckley 1985) and Krovetz stemming (Krovetz 1993). For each document in the test collection, we look up its record in our database and append any citation contexts it has there as text strings to the XML document before indexing. We then build one index from the XML documents alone and another from each of the document-plus-citation-context representations. In order to investigate the effect of weighting citation terms differently relative to document terms, we had two options in our experimental set-up. The first is to create a version of each document for each citation weight, where the citation context strings are added in duplicate to the base XML document to achieve the desired weight. Separate indexes are then built from each of these weighted document collections.

The second option is to use weighting operators in the Indri query language to weight terms according to which part of the document they occur in, i.e., which *field*. In this method, the weight is applied to the citation terms at query time, as part of the query-document match cal-

culation, rather than requiring duplication of the terms in the indexed document. However, this method can only be used with the Indri retrieval model, since Indri's query language and retrieval model are integrated components; we cannot use weighted queries to investigate the effects of citation term weighting on other models' effectiveness, even those other models implemented in Lemur. There is a field-weighted version of the Okapi BM25 retrieval function, BM25F (Robertson, Zaragoza & Taylor 2004), but this is not implemented in Lemur. Therefore, we opted for the term duplication method, rather than restrict our investigation to a single retrieval model. The method is resource-hungry, however, and we investigate a limited number of weights in this way.

Further to the practical differences, the two weighting methods are not equivalent in terms of document scoring and ranking, for multiple reasons. Firstly, the weighted query method calculates term counts and smoothing parameters calculated across individual fields, rather than across whole documents, as in the case of unweighted queries. Thus, the relative scores and the ranking produced by a weighted query where the fields are weighted equally and the ranking produced by its unweighted counterpart on the same index will not necessarily be the same. To illustrate, consider the example in Figure 6.1 of a single-term query `cinnamomeous`, run against a document collection including documents A and B[1]. The query term appears three times in document A, which has 5042 terms in total, each time in field F1; the term appears once in document B, which has 3580 terms, in field F2. We omit the smoothing calculation details but show enough of the calculation to illustrate how the difference between the term count for the collection (in the unweighted query) versus the term counts for the individual fields (in the weighted query) results in different scores for the two documents and a different ranking. Intuitively, the document scores and rankings should be the same for both queries; that they should be the same was one of the factors that led to the design of BM25F (Robertson et al. 2004).

Secondly, in the term duplication method, the term counts for a given term will be different in each index, as it is altered by the citation 'weight': there will be an additional occurrence of that term in the index for every duplicate citation term that is added. This is not the case in the weighted query method, where each citation term is added exactly once to the index. Thus, the term duplication method will not give equivalent results to either the Indri weighted queries or to weighted queries where the term counts are calculated across whole documents. The differences between these weighting methods opens the door for comparative experimentation between them but this is outside the scope of this thesis.

Our queries are stopped and stemmed in the same way as the documents. We use Lemur's implementations of the following retrieval models with standard parameters to test our method:

Cosine The cosine similarity model

Okapi The Okapi BM25 retrieval function

Indri The Indri structured query retrieval model

KL The Kullback-Leibler (KL) divergence language model based retrieval method

KL FB The KL divergence method with relevance feedback

[1]The example presented here is based on a discussion of a real example with Indri developers. Many thanks to David Fisher for his calculations.

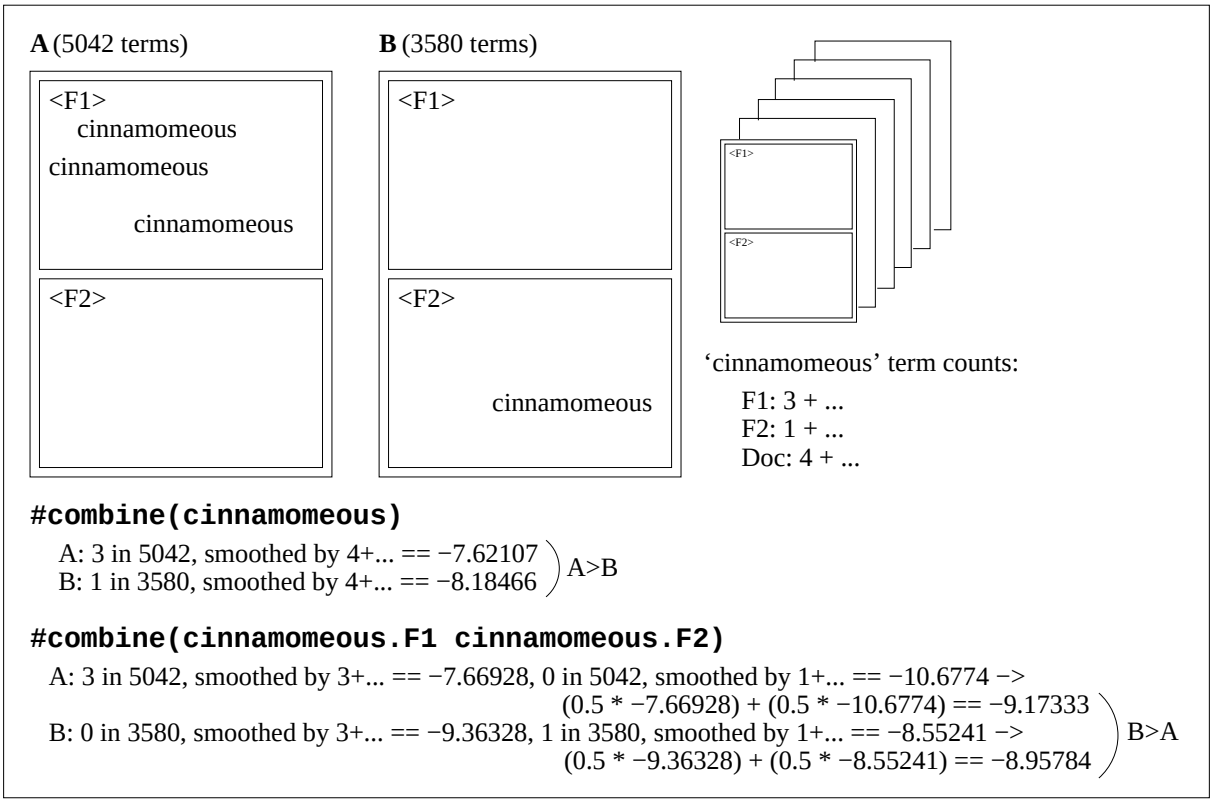


Figure 6.1: Example calculations of unweighted and uniformly weighted Indri query scores.

For KL FB, we use each query's entire set of relevance judgements for feedback. Using the same judgements for feedback and evaluation is an unrealistic experiment. However, it affords us some idea of the relative effectiveness that may be achieved with more sophisticated retrieval methods than the basic models. We do not report results using Okapi with relevance feedback, since the Lemur documentation notes a suspected bug in the Okapi feedback implementation. In each run, 100 documents were retrieved per query. Nowadays, it is typical for greater numbers of documents to be retrieved in evaluations, e.g., 1000 in TREC tracks. However, 100 documents is already far greater than the number of judged documents for any query in our test collection; the top 100 documents should encapsulate any important effects of our citation experiments on the rankings. For evaluation, we use the TREC evaluation software, `trec_eval`[2], and report a number of standard evaluation measures, discussed in the following section.

6.2 Evaluation measures

Intuitively, an IR system is successful when it retrieves all documents that are relevant to a given query and no irrelevant documents. Most IR evaluation measures are, therefore, based on *precision* and *recall*. Precision is a measure of system ability to present only relevant documents: the proportion of retrieved documents that are relevant. Recall is a measure of system ability to present all relevant documents: the proportion of the total relevant documents that are retrieved. These basic concepts are set-based, i.e., they evaluate unordered sets of retrieved documents.

[2] `http://trec.nist.gov/trec_eval/trec_eval.8.1.tar.gz`

Nowadays, it is typical to use more comprehensive measures that take into account the order in which systems present documents to the user, i.e., the document ranking. Usually, scores will be presented as an average across queries. In this section, we describe a number of common evaluation measures; specifically, we describe the set of measures calculated by `trec_eval` that we consider for the evaluation of the experiments presented in this thesis.

Precision at 5 documents

Precision at 5, henceforth P@5, is a basic precision score calculated after the first five retrieved documents. So, for example, a system which returns only one relevant document in the top five will receive a score of 0.2. Five documents is the 'shortest' of a series of standard document ranking cut-offs at which precision scores are commonly calculated. The advantage of these measures are that they are straightforward to interpret and are directly comparable between queries, since they are calculated over the same number of documents for every query. On the other hand, they do not take into account the ranks at which relevant documents are retrieved within the given portion of the ranking, i.e., a system which retrieves one relevant document at the top rank will receive the same score as a system which retrieves one relevant document but at rank 5. Also, these measures are not sensitive to the differing total numbers of relevant documents that queries have. For some queries, this will be less than the document cut-off and, in these cases, it will be impossible to achieve a 'perfect' score of 1.0 by retrieving all of the relevant documents at the top of the ranking. For all queries, relevant documents retrieved outside of the prescribed portion of the ranking will not be taken into account; this will particularly affect those queries which have more relevant documents than the cut-off.

R-precision

R-precision is the precision score after the first R retrieved documents, where R is the number of relevant documents for that query. This measure is designed to overcome the problem of differing numbers of relevant documents that precision scores at fixed cut-offs suffer from. However, R-precision still does not take the ranks of retrieved documents into account and, thus, like P@5, does not distinguish between systems that retrieve the same number of relevant documents in different positions within the top R.

Mean Average Precision

Mean average precision (MAP) is the most commonly used of all IR measures. It is defined as the arithmetic mean (over queries) of average precisions, where average precision is the mean of the precision scores after each relevant document is retrieved:

$$MAP = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{R_i}\sum_{j=1}^{R_i} P@reldoc_j$$

Q : the set of queries

R_i : the number of judged relevant documents for query i

$P@reldoc_j$: precision after j th relevant document

By its definition, MAP contains both precision and recall oriented aspects, and is sensitive to the entire ranking. However, it is less readily interpretable than simpler measures, such as

P@5. For instance, considering the simplest case of a single query, whereas a P@5 score of 0.2 always means that exactly one relevant document was retrieved in the top five, a MAP score of 0.2 could mean that the query has only relevant document and it was returned at rank 5 or that it has two relevant documents, one of which was returned at rank 5 and the other at 10 etc.

bpref

Buckley & Voorhees (2004) introduced bpref, a measure designed for use with incomplete sets of relevance judgements. The measure takes into account only those retrieved documents for which explicit relevance judgements have been made, rather than assuming that unjudged documents are irrelevant. It computes a preference relation of whether judged relevant documents are retrieved ahead of judged irrelevant documents:

$$bpref = \frac{1}{R}\sum_{r}(1 - \frac{|n\_ranked\_higher\_than\_r|}{min(R,N)}$$

R : the number of judged relevant documents
N : the number of judged irrelevant documents
r : a relevant retrieved document
n : a member of the first R irrelevant retrieved documents

bpref can be thought of as the inverse of the fraction of judged irrelevant documents that are retrieved before relevant ones, i.e., the higher the value of bpref, the better the system has performed. When judgements are complete, system rankings generated by bpref scores are very similar to those generated by MAP scores, i.e., the rank correlation is high, as measured by Kendall's τ. With incomplete judgements, on the other hand, bpref system rankings still correlate highly with the ranking from the complete judgement set, whereas MAP rankings do not.

Relevant documents retrieved

A crude way to gauge the relative success of different systems is to simply count how many relevant documents they each retrieve for the same query or queries. This measure does not take into account either the total number of relevant documents for a query or the ranks at which relative documents are retrieved. Henceforth, we will denote the number of relevant documents retrieved by the abbreviation #RR.

6.3 Sanity check experiments and ranking notation

We first present some results of some baseline retrieval runs, to demonstrate the validity of our test collection as an evaluation tool. We also introduce some notation. Table 6.1 shows how the five retrieval models are ranked according to each of the five evaluation measures, running the queries against the basic index with no citation terms. Our notation is as follows: a) models are ranked by absolute effectiveness values, in ascending order from left to right; b) $\ll$ denotes the difference between a pair of models is significant for p$\leq$0.01, $<$ denotes significance for p$\leq$0.05 and $\approx$ denotes statistical insignificance; and c) we assume a subsumptive, transitive significance relation in the ranking, i.e., if $A \ll B \approx C$ (or $A \ll B < C$) then $A \ll C$ etc. We note exceptions to this general ranking in brackets in the rightmost column; these anomalies are independent of

Measure	Model ranking (and anomalies)
MAP	Okapi $\ll$ Cosine $\approx$ Indri $\ll$ KL $\ll$ KL FB
P@5	Okapi $<$ Cosine $\ll$ Indri $\approx$ KL $<$ KL FB (Indri $\ll$ KL FB)
R-P	Okapi $<$ Cosine $\ll$ Indri $\approx$ KL $\ll$ KL FB
bpref	Okapi $\ll$ Cosine $<$ Indri $\ll$ KL $\ll$ KL FB
#RR	Okapi $\ll$ Cosine $\ll$ Indri $\ll$ KL $\ll$ KL FB

Table 6.1: Baseline model rankings for different evaluation measures.

each other and affect only the listed pair of contexts so, e.g., if $A \approx B \approx C \approx D$ has the exception $A < C$, it does not follow that $A < D$. We use Student's t-test to test for statistical significance of differences between models' average effectiveness scores. The t-test assumes that the individual query effectiveness scores are distributed normally, which is generally untrue. Nevertheless, Sanderson & Zobel (2005) found that the t-test is highly reliable for significance testing in IR.

For all five evaluation measures, Okapi is ranked lowest, followed by Cosine, then Indri, KL and, finally, KL FB is ranked highest. In all but two cases, the differences between models are statistically significant. The ranking produced by our test collection is stable across performance measures and the differences between models are significant. This is largely the ranking that might be expected from this set of models: language modelling methods like KL and Indri typically outperform both Okapi, in Lemur (e.g., Ogilvie & Callan 2001, Bennett, Scholer & Uitdenbogerd 2008) and other systems (e.g., Garcia, Lester, Scholer & Shokouhi 2006), and vector space models like Cosine (e.g., Taghva, Coombs, Pareda & Nartker 2004, Aslam, Pavlu & Rei 2006); incorporating relevance feedback into a model generally increases effectivess (e.g., Harman 1992). The marked underperformance of Okapi to Cosine is a surprising result. This may be a product of the unusual nature of the queries and documents, compared to the test collections that have typically been used in IR evaluations; perhaps it is a characteristic of specific, scientific queries such as ours that vector space models will perform uncharacteristically well, outperforming probabilistic models. Investigating this issue, however, would seem to require an 'equivalent' test collection to compare against, which as yet does not exist.

These rankings are nevertheless a satisfactory indication that our collection is successfully distinguishing between the different models' performance: the collection is a useful tool for IR evaluation, despite the fact that the relevance judgements are almost certainly incomplete, even after Phase Two.

6.4 Basic citation experiments

In this section, we present a basic comparison between using the document terms alone and using the document plus citation terms. The core hypothesis of this thesis was that an existing document representation would be enhanced by adding to it terms from citations to the document; this is the fundamental experiment that will test our hypothesis. We use the default citation context of the sentence that contains the citation; we weight citation and document terms equally. Table 6.2 summarises the results. In each row, we compare the retrieval effectiveness of a given retrieval model on the index without citation terms to its effectiveness on the index with citation terms. For each evaluation measure, we present the value of that measure on the with-citation index in the left of the column and the difference from the corresponding without-citation value on the right; a positive difference indicates that effectiveness

Retrieval model	Evaluation measure									
	MAP		P@5		R–precision		bpref		#RR	
Okapi	.096	+.002	.132	+.002	.113	+.004	.245	**+.009**	270	**+10**
Cosine	.142	+.002	.195	+.005	.152	+.007	.323	**+.014**	439	**+30**
Indri	.176	**+.013**	.283	**+.019**	.212	+.008	.392	**+.025**	502	**+38**
KL	.189	**+.011**	.300	+.015	.217	**+.011**	.395	+.006	535	**+29**
KL FB	.221	**+.012**	.324	**+.020**	.253	+.011	.447	+.011	586	**+32**

Table 6.2: Retrieval effectiveness with citations versus without citations.

was higher with citations. We t-test for statistical significance of with- versus without-citation effectiveness; differences highlighted in bold are significant for $p \leq 0.05$ and those underlined for $p \leq 0.01$.

Effectiveness is uniformly higher with citations than without, for all models, for all measures. More than half the differences are statistically significant. Notably, MAP increases by as much as 7.4% (for Indri), P@5 by 6.7% (for Indri), R-precision by 5.1% (for KL) and bpref by 6.4% (for Indri), and a significant number of previously unretrieved relevant documents are discovered by all models when citation terms are indexed. These results support this hypothesis: adding citation terms to the document representation does, indeed, improve retrieval effectiveness. We now go on to investigate variations on the basic method and what effect these have on retrieval effectiveness.

6.5 Weighting experiments

Here, we investigate the effect of weighting citation terms higher relative to document terms. Table 6.3 gives the results for weights of 1 to 5 and also 11 and 35. These two larger weights were selected since, in various TREC Web retrieval tasks, optimal Okapi effectiveness has been achieved by weighting anchor text 11.5[3] and 35 times higher than web page body text, in Topic Distillation and Named Page Finding, respectively (Zaragoza, Craswell, Taylor, Saria & Robertson 2004).

Though the number of citation weights investigated is limited, for practical reasons, the results nevertheless allow some interesting observations to be made. Firstly, the general trend is for effectiveness to increase as citation terms are weighted higher. Most increases are significant for Indri, KL and KL FB: 80% across all measures, including 95% of MAP increases, 90% of #RR increases and 76% of bpref increases. Cosine and Okapi show the smallest and least significant effectiveness increases. The results for Cosine, in particular, do not exhibit the trend of increasing effectiveness with increasing citation term weight as clearly as the other models.

Secondly, for most models, the increase in effectiveness has clearly diminished by the time a weight of 35 is reached. For some models, effectiveness according to some measures even drops (insignificantly) below that of the without-citations baseline: Okapi MAP and R-precision drop by 0.005 (5.7%) and 0.007 (6.9%), respectively; Cosine MAP, P@5 and R-precision drop by 0.009 (6.9%), 0.002 (1.1%) and 0.003 (2.1%); KL P@5 drops by 0.005 (1.8%). For Okapi and Indri, the optimal weight for citation terms might even be in the range 1–5, as the values of some measures appear to have plateaued and begun to decrease again within this window.

[3]With our simple weighting implementation, a weight of 11.5 would be achieved by duplicating the document terms twice and the citation terms 22 times; thus, we experimented with an integer weight of 11 instead.

Retrieval model	W	Evaluation measure									
		MAP		P@5		R–precision		bpref		#RR	
Okapi	1	.096	+.002	.132	+.002	.113	+.004	.245	**+.009**	270	**+10**
	2	.096	+.003	.146	+.017	.116	+.007	.247	**+.012**	276	**+16**
	3	.097	+.004	.154	**+.024**	.119	+.011	.247	**+.012**	277	**+17**
	4	.097	+.004	.156	**+.027**	.119	+.010	.247	**+.012**	278	**+18**
	5	.098	+.004	.151	**+.022**	.121	+.012	.247	+.011	277	**+17**
	11	.094	+.001	.146	+.017	.111	+.002	.246	+.010	279	**+19**
	35	.088	–.005	.139	+.010	.102	–.007	.242	+.007	273	+13
Cosine	1	.142	+.002	.195	+.005	.152	+.007	.323	**+.014**	439	**+30**
	2	.145	+.004	.200	+.010	.153	+.008	.325	**+.016**	442	**+33**
	3	.142	+.002	.202	+.012	.154	+.009	.328	**+.019**	447	**+38**
	4	.144	+.004	.202	+.012	.157	+.011	.328	**+.020**	446	**+37**
	5	.143	+.002	.205	+.015	.156	+.011	.319	+.010	450	**+41**
	11	.141	+.001	.210	+.020	.153	+.008	.328	+.020	458	**+49**
	35	.131	–.009	.188	–.002	.142	–.003	.320	+.011	427	+18
Indri	1	.176	**+.013**	.283	**+.019**	.212	+.008	.392	**+.025**	502	**+38**
	2	.190	**+.027**	.312	**+.049**	.223	+.020	.398	**+.031**	524	**+60**
	3	.193	**+.030**	.320	**+.056**	.226	+.023	.400	**+.033**	527	**+63**
	4	.197	**+.034**	.317	**+.054**	.234	**+.031**	.402	**+.035**	529	**+65**
	5	.198	**+.035**	.307	**+.044**	.235	**+.031**	.403	**+.035**	526	**+62**
	11	.196	**+.033**	.324	**+.061**	.231	+.027	.402	**+.034**	519	**+55**
	35	.183	**+.020**	.315	**+.051**	.218	+.015	.392	**+.024**	487	+23
KL	1	.189	**+.011**	.300	+.015	.217	**+.011**	.395	+.006	535	**+29**
	2	.192	**+.014**	.293	+.007	.222	**+.017**	.401	+.012	549	**+43**
	3	.196	**+.017**	.290	+.005	.227	**+.021**	.404	+.015	552	**+46**
	4	.196	**+.018**	.293	+.007	.230	**+.024**	.412	**+.022**	560	**+54**
	5	.198	**+.020**	.293	+.007	.234	**+.028**	.415	**+.026**	561	**+55**
	11	.198	**+.020**	.290	+.005	.237	**+.031**	.414	**+.025**	548	**+42**
	35	.189	+.011	.281	–.005	.230	+.024	.409	+.020	531	+25
KL FB	1	.221	**+.012**	.324	**+.020**	.253	+.011	.447	+.011	586	**+32**
	2	.268	**+.059**	.346	**+.041**	.283	**+.041**	.509	**+.073**	624	**+70**
	3	.271	**+.062**	.349	**+.044**	.293	**+.050**	.517	**+.081**	629	**+75**
	4	.276	**+.067**	.359	**+.054**	.299	**+.057**	.523	**+.087**	634	**+80**
	5	.276	**+.067**	.363	**+.059**	.304	**+.061**	.527	**+.091**	634	**+80**
	11	.284	**+.075**	.378	**+.073**	.315	**+.072**	.533	**+.097**	633	**+79**
	35	.278	**+.069**	.363	**+.059**	.312	**+.070**	.537	**+.101**	626	**+72**

Table 6.3: Retrieval effectiveness with weighted citations versus without citations.

This contrasts with the much higher anchor text weights found to be effective with Okapi on Web tasks. The results for KL suggest its optimal weight will be somewhere between 5 and 11; for KL FB, it may be even greater than 35. In general, further investigation is required to more accurately discover the optimal weighting of citation terms to document terms. Nevertheless, these results suggest that weighting citation terms higher relative to document terms generally improves retrieval effectiveness.

Figure 6.2 presents a selection of our results graphically, to show the general trends: (a) shows how bpref changes with increasing citation term weight for each of the retrieval models, while (b) shows how each of the effectiveness measures change for the KL model. The complete set of plots is given in Appendix D.

6.6 Context experiments

Finally, we investigate how retrieval effectiveness is affected by exactly where from around citations terms are taken, i.e., what the definition of the citation context is. There is no anchor text in scientific papers, unlike in web pages, where there are HTML tags to delimit the text associated with a link. Identifying which words are associated with a citation is an interesting, complex problem, which has been discussed in depth (O'Connor 1982, Ritchie et al. 2006*b*). These independent case studies, on different domains, have each provided evidence to suggest

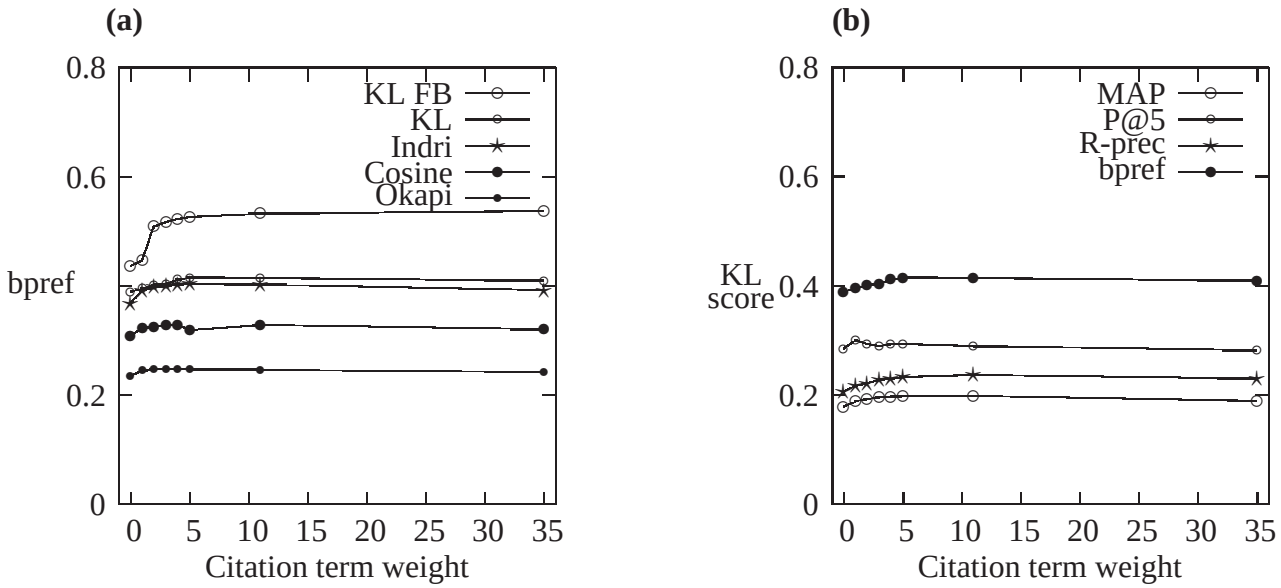


Figure 6.2: Retrieval effectiveness with increasing citation term weight: **(a)** all models' bpref scores and **(b)** KL's scores for all measures.

that computational linguistics (CL) techniques may be useful for more accurately locating these citation terms, e.g., using topic shift detection to indicate when the text has moved on from the subject of the citation. We do not attempt a comprehensive exploration of CL techniques here: that is beyond the scope of our investigation. We do, however, compare a number of alternative contexts that range in size and complexity, some of which are very basically linguistic in nature, e.g., use knowledge of sentence boundaries. Specifically, we define a citation context in the following nine different ways. The bracketed numbers indicate the average number of words in the left and right portions of these contexts, respectively, in the corresponding index.

`none` No citation context. [0,0]

`1sent` The sentence containing the citation. [13.6,10.6]

`3sent` The sentence containing the citation plus one sentence immediately to the left and right. [23.3,23.0]

`1sentupto` The sentence containing the citation, truncated at the next citations to the left and right. [9.8,8.1]

`3sentupto` The `3sent` context, truncated at the next citations to the left and right. [13.6,17.0]

`win50` A window of up to 50 words on each side of the citation. [49.0,49.4]

`win75` A window of up to 75 words on each side of the citation. [70.2,70.2]

`win100` A window of up to 100 words on each side of the citation. [84.7,84.3]

`full` The entire citing paper.

The average number of terms in the `win100` contexts is notably different from the maximum allowed by its design. This is because the database field from which the window contexts are taken consisted of a fixed number of sentences around the citation and, for some citations, this

group of sentences will not include as many as 100 words on either side. This is an unfortunate error in the database design, which took place before the scope of these experiments was realised.

The rationale behind this range of contexts is as follows. We start with the basic assumption that the words that are used to describe the cited paper will occur close to the citation and that, further away, words are less likely to be about the paper. We also suppose that the citing author will have constructed their text according to grammatical and rhetorical conventions, so sentence boundaries should demarcate logical units of related text. Intuitively, then, the sentence that contains the citation is a good first approximation of the extent of its descriptive terms and, therefore, this is our default context, called `1sent`. We also consider whether descriptive terms might be found outside the citing sentence, and investigate the context of the three sentences immediately around the citation, `3sent`. The `1sentupto` and `3sentupto` contexts are devised to investigate whether neighbouring citations might be an effective indicator of the end of the text associated with the original citation. Then, in contrast to these linguistically motivated contexts, we look at some fixed window contexts, to compare the effectiveness of simpler methods of taking terms from around citations, `win50`, `win75` and `win100`. Finally, we include in the comparison the two extreme cases of adding no citation context, `none`, and adding the entire citing paper as the citation context, `full`.

In each row of Table 6.4, we compare the effectiveness of a given retrieval model with the default context, `1sent`, to its effectiveness with another context. We compare with `1sent`, as opposed to `none`, since we have already shown that adding `1sent` to the base document representation results in an improvement in effectiveness; we hence treat `1sent` as the baseline for this comparison of citation contexts[4]. As usual, the difference between `1sent` and the other context is given in the right hand side of each column; positive differences indicate that `1sent`'s effectiveness was higher and differences highlighted in bold are significant for $p \leq 0.05$ and underlined differences are significant for $p \leq 0.01$. There are 200 equivalent pairwise comparisons between all contexts and this generates a further seven tables like Table 6.4. For brevity, we summarise this information in the context rankings given in Tables 6.5 and 6.6, using the ranking notation introduced in Section 6.3.

A first notable outcome from these results is that they confirm that adding citation terms improves retrieval effectiveness. Effectiveness is generally better on the indexes with citation terms than on the `none` index: `none` is ranked lowest in 14 of the 25 rankings (i.e., combinations of retrieval model and evaluation measure). In 185 of 200 pairwise comparisons with other contexts, `none` is ranked lower and, in 107 of these, the difference is significant. In the 11 rankings where `none` is not the lowest ranked context, the lowest is `full`; in five of these cases the difference between `none` and `full` is significant, all with Okapi. In the remaining 10 of the 15 comparisons where `none` is ranked above another context, the difference is insignificant.

Looking in further detail at the retrieval effectiveness achieved by adding the entire citing paper, `full` is ranked higher and lower than `none` in almost equal measure: higher in 14 rankings, only four of whose differences are significant, and lower in 11 rankings, five of whose differences are significant. The majority of differences between `full` and `none` are insignificant. In contrast to the five rankings where `full` is ranked lowest overall, it is ranked highest

[4]The apparent discrepancies between the values in Table 6.4 and those in the previous tables are due to rounding. For instance, Table 6.4 states that Okapi's MAP with `none` is 0.093 (rounded from 0.0931) where this is 0.002 smaller than the `1sent` value, implying a value of 0.095. However, the exact `1sent` value is 0.0955 and Table 6.2 accordingly gives the value as 0.096. The exact difference of 0.0024 is rounded to 0.002.

Context	Retrieval model	Evaluation measure									
		MAP		P@5		R–precision		bpref		#RR	
none	Okapi	.093	(–.002)	.129	(–.002)	.109	(–.004)	.235	**(–.009)**	260	**(–10)**
	Cosine	.140	(–.002)	.190	(–.005)	.145	(–.007)	.309	**(–.014)**	409	**(–30)**
	Indri	.163	**(–.013)**	.263	**(–.019)**	.204	(–.008)	.367	**(–.025)**	464	**(–38)**
	KL	.178	**(–.011)**	.285	(–.015)	.206	**(–.011)**	.389	(–.006)	506	**(–29)**
	KL FB	.209	**(–.012)**	.305	**(–.020)**	.242	(–.011)	.436	(–.011)	554	**(–32)**
3sent	Okapi	.096	(+.000)	.134	(+.002)	.107	(–.006)	.241	(–.004)	272	(+2)
	Cosine	.152	**(+.010)**	.200	(+.005)	.161	**(+.009)**	.327	(+.004)	452	**(+13)**
	Indri	.191	**(+.015)**	.302	**(+.020)**	.221	(+.009)	.406	**(+.014)**	527	**(+25)**
	KL	.199	**(+.010)**	.307	(+.007)	.225	**(+.008)**	.405	**(+.010)**	553	**(+18)**
	KL FB	.232	**(+.010)**	.332	(+.007)	.251	(–.003)	.455	(+.009)	600	**(+14)**
1sentupto	Okapi	.094	(–.001)	.139	(+.007)	.114	(+.001)	.241	(–.004)	264	**(–6)**
	Cosine	.142	(–.001)	.195	(+.000)	.150	(–.002)	.315	**(–.008)**	424	**(–15)**
	Indri	.171	**(–.005)**	.273	(–.010)	.208	(–.004)	.375	**(–.017)**	485	**(–17)**
	KL	.186	(–.003)	.298	(–.002)	.209	**(–.008)**	.395	(–.001)	524	**(–11)**
	KL FB	.218	(–.004)	.320	(–.005)	.244	(–.009)	.445	(–.002)	578	**(–8)**
3sentupto	Okapi	.097	(+.002)	.139	(+.007)	.115	(+.001)	.241	(–.004)	262	**(–8)**
	Cosine	.146	(+.004)	.198	(+.003)	.153	(+.001)	.324	(+.001)	440	(+1)
	Indri	.179	(+.003)	.281	(–.002)	.214	(+.002)	.392	(+.000)	508	(+6)
	KL	.192	(+.003)	.300	(+.000)	.218	(+.001)	.399	(+.004)	534	(–1)
	KL FB	.222	(+.001)	.317	(–.007)	.245	(–.008)	.446	(–.001)	584	(–2)
win50	Okapi	.094	(–.002)	.134	(+.002)	.102	(–.011)	.237	(–.008)	266	(–4)
	Cosine	.155	**(+.012)**	.207	(+.012)	.163	(+.010)	.333	(+.009)	456	**(+17)**
	Indri	.187	**(+.011)**	.290	(+.007)	.217	(+.005)	.400	(+.007)	522	**(+20)**
	KL	.194	(+.005)	.310	(+.010)	.231	**(+.014)**	.403	(+.007)	546	(+11)
	KL FB	.227	(+.006)	.324	(+.000)	.252	(–.001)	.460	(+.013)	602	**(+16)**
win75	Okapi	.097	(+.002)	.139	(+.007)	.103	(–.010)	.238	(–.007)	268	(–2)
	Cosine	.155	**(+.013)**	.210	(+.015)	.171	**(+.018)**	.332	(+.009)	458	**(+19)**
	Indri	.197	**(+.021)**	.329	**(+.046)**	.227	(+.015)	.404	(+.012)	544	**(+42)**
	KL	.197	(+.008)	.305	(+.005)	.221	(+.005)	.408	(+.013)	562	**(+27)**
	KL FB	.266	**(+.044)**	.368	**(+.044)**	.285	**(+.032)**	.494	**(+.047)**	632	**(+46)**
win100	Okapi	.100	(+.004)	.149	(+.017)	.103	(–.011)	.247	(+.002)	273	(+3)
	Cosine	.156	**(+.014)**	.212	(+.017)	.171	**(+.019)**	.331	(+.008)	460	**(+21)**
	Indri	.204	**(+.028)**	.337	**(+.054)**	.235	(+.023)	.408	(+.016)	548	**(+46)**
	KL	.202	(+.012)	.307	(+.007)	.225	(+.009)	.409	(+.014)	558	**(+23)**
	KL FB	.277	**(+.056)**	.381	**(+.056)**	.295	**(+.042)**	.520	**(+.073)**	628	**(+42)**
full	Okapi	.048	**(–.048)**	.071	**(–.061)**	.056	**(–.057)**	.174	**(–.071)**	178	**(–92)**
	Cosine	.149	(+.007)	.205	(+.010)	.162	(+.010)	.364	**(+.041)**	482	**(+43)**
	Indri	.176	(+.000)	.283	(+.000)	.216	(+.004)	.403	(+.011)	529	(+27)
	KL	.166	(–.023)	.244	**(–.056)**	.187	(–.029)	.411	(+.016)	533	(–2)
	KL FB	.200	(–.021)	.290	(–.034)	.220	(–.033)	.460	(+.013)	595	(+9)

Table 6.4: Retrieval effectiveness with various citation contexts versus 1sent.

three times; in one of these, the difference between full and every other context is significant. Thus, it appears that the effect of full is unpredictable and that, overall, there is no advantage to additionally indexing the citing paper over indexing the cited paper alone. Now comparing full with the restricted citation contexts, full appears to be less effective. The difference between full and 1sent is marginal: full is ranked below 1sent in 12 rankings, of which seven differences are significant, compared to 11 rankings where it is ranked higher, of which only two differences are significant. The difference is more marked between full and the window contexts: full is ranked above, e.g., win50 in five rankings, in only one of which the difference is significant, compared to 19 rankings where full is ranked lower, in 10 of which the difference is significant. The gap is even wider between full and the longer window contexts. Using terms from a limited context around citations is, thus, more effective than using the entire citing paper, reinforcing our second observation that indexing the entire citing paper is not a worthwhile method. This is not a surprising conclusion: intuitively, the vast majority of the words in the citing paper will not refer to the cited paper and will probably not be appropriate index terms for it. Moreover, this large number of 'bad' index terms could potentially

Model	Measure	Context ranking	Ranking anomalies
Cosine	MAP	none ≈ 1sentupto ≪ 1sent ≈ 3sentupto ≈ full ≪ 3sent ≈ win50 ≈ win75 ≈ win100	none≈1sent, none<3sentupto, none≈full, 1sentupto≈1sent, 1sentupto≈3sentupto, 1sentupto≈full, 3sentupto<win100, full≈3sent, full≈win50, full≈win75, full≈win100
	P@5	none ≈ 1sent ≈ 1sentupto ≈ 3sentupto ≈ 3sent ≈ full ≈ win50 ≈ win75 ≈ win100	
	R-prec	none ≪ 1sentupto ≈ 1sent ≈ 3sentupto ≈ 3sent ≈ full ≈ win50 ≪ win100 ≈ win75	none≈1sentupto, none<1sent, none<3sentupto, none≈full, 1sentupto≪3sent, 1sentupto<win50, 1sent≪3sent, 1sent<win50, 3sentupto<3sent, 3sent≈win100, 3sent<win75, full≈win100, full≈win75, win50≈win100, win50<win75
	bpref	none ≪ 1sentupto < 1sent ≈ 3sentupto ≈ 3sent ≈ win100 ≈ win75 ≈ win50 ≪ full	none<1sentupto, none<3sent, none<win100, 1sentupto≪3sentupto, 1sentupto≪win50
	#RR	none ≪ 1sentupto ≪ 1sent ≈ 3sentupto ≪ 3sent ≈ win50 ≈ win75 ≈ win100 ≈ full	1sent<3sent, 1sent<win100, 3sentupto<3sent, 3sentupto<win100, 3sent<full
Indri	MAP	none ≪ 1sentupto ≈ full ≈ 1sent ≈ 3sentupto ≈ win50 ≪ 3sent ≈ win75 ≪ win100	none≈full, 1sentupto≪1sent, 1sentupto≪3sentupto, 1sentupto≪win50, full≈3sent, full≈win75, full<win100, 1sent≪win50, 3sentupto≪3sent, win50≈3sent, win50<win75, 3sent<win100
	P@5	none ≈ 1sentupto ≈ 3sentupto ≈ full ≈ 1sent ≈ win50 ≪ 3sent < win75 ≈ win100	none<1sent, none<win50, full≈3sent, full<win75, full<win100, win50≈3sent, 3sent<win100
	R-prec	none ≈ 1sentupto ≈ 1sent ≈ 3sentupto ≈ full ≈ win50 ≈ 3sent ≈ win75 ≈ win100	none<3sentupto, none<3sent, none<win100, 1sentupto≪3sent, 1sentupto<win100, 1sent<3sent, 1sent<win100
	bpref	none < 1sentupto ≪ 1sent ≈ 3sentupto ≈ win50 ≈ full ≈ win75 ≈ 3sent ≈ win100	none≈full, 1sentupto≈full, 1sent≪3sent, 3sentupto≪3sent
	#RR	none ≪ 1sentupto ≪ 1sent ≈ 3sentupto ≈ win50 ≪ 3sent ≈ full ≈ win75 ≈ win100	1sentupto<full, 1sent<win50, 1sent≈full, 3sentupto≈full, win50≈3sent, win50≈full, win50<win75, 3sent<win100

Table 6.5: Condensed rankings of citation contexts by retrieval effectiveness (for Cosine and Indri).

Model	Measure	Context ranking	Ranking anomalies
KL	MAP	full ≈ none ≪ 1sentupto ≈ 1sent ≈ 3sentupto ≈ win50 ≈ win75 ≈ 3sent ≈ win100	full≈1sentupto, full≈1sent, full<3sentupto, full<win50, none<win75, 1sentupto≪3sentupto, 1sentupto<3sent, 1sentupto<win100, 1sent<3sent
	P@5	full ≪ none ≈ 1sentupto ≈ 1sent ≈ 3sentupto ≈ win75 ≈ win100 ≈ 3sent ≈ win50	full≈none, full<1sentupto, full<1sent, full<3sentupto, none<1sentupto, none<win50
	R-prec	full ≈ none ≈ 1sentupto ≪ 1sent ≈ 3sentupto ≈ win75 ≈ win100 ≈ 3sent ≈ win50	full≈1sent, full≈3sentupto, full<win75, full<3sent, none<3sentupto, none≈win75, none<win100, 1sentupto<3sentupto, 1sentupto≈win75, 1sentupto≈win100, 1sent≪3sent, 1sent≪win50
	bpref	none ≈ 1sent ≈ 1sentupto ≈ 3sentupto ≈ win50 ≈ 3sent ≈ win75 ≈ win100 ≈ full	1sent<3sent
	#RR	none ≪ 1sentupto ≪ full ≈ 3sentupto ≈ 1sent ≈ win50 ≈ 3sent < win100 ≈ win75	none≈full, 1sentupto≈full, 1sentupto<3sentupto, 1sentupto<1sent, 3sentupto≪3sent, 3sentupto≪win75, 1sent≪3sent, 1sent≪win75, win50≈win100, 3sent≈win100, 3sent≈win75
KL FB	MAP	full ≈ none ≪ 1sentupto ≈ 1sent ≈ 3sentupto ≈ win50 ≪ 3sent ≪ win75 < win100	full≈1sentupto, full≈1sent, full≈3sentupto, full<win50, full<3sent, win50≈3sent
	P@5	full ≈ none ≈ 3sentupto ≈ 1sentupto ≈ win50 ≈ 1sent ≈ 3sent ≪ win75 ≈ win100	none<1sentupto, none<win50, none<1sent, none≪3sent, 3sentupto<3sent
	R-prec	full ≈ none ≈ 1sentupto ≈ 3sentupto ≈ 3sent ≈ win50 ≈ 1sent ≪ win75 ≈ win100	full<win50, full<1sent, none<1sent, 1sentupto<1sent
	bpref	none ≈ 1sentupto ≈ 3sentupto ≈ 1sent ≈ 3sent ≈ full ≈ win50 ≪ win75 ≈ win100	none<win50, 1sentupto<win50, full≈win75,
	#RR	none ≪ 1sentupto ≈ 3sentupto ≈ 1sent ≈ full ≪ 3sent ≈ win50 ≪ win100 ≈ win75	none<full, 1sentupto<1sent, 1sent<3sent, 1sent<win50, full≈3sent, full≈win50
Okapi	MAP	full ≪ none ≈ win50 ≈ 1sentupto ≈ 3sent ≈ 1sent ≈ 3sentupto ≈ win75 ≈ win100	win50<win100
	P@5	full ≪ none ≈ 1sent ≈ win50 ≈ 3sent ≈ 1sentupto ≈ 3sentupto ≈ win75 ≈ win100	
	R-prec	full ≪ win50 ≈ win100 ≈ win75 ≈ 3sent ≈ none ≈ 1sent ≈ 1sentupto ≈ 3sentupto	
	bpref	full ≪ none ≈ win50 ≈ win75 ≈ 3sentupto ≈ 3sent ≈ 1sentupto ≈ 1sent ≈ win100	full<none, none<1sent
	#RR	full ≪ none ≈ 3sentupto ≈ 1sentupto ≈ win50 ≈ win75 ≈ 1sent ≈ 3sent ≈ win100	none<1sent, 3sentupto<1sent, 1sentupto<1sent

Table 6.6: Condensed rankings of citation contexts by retrieval effectiveness (for KL, KL FB and Okapi).

drown out the ‘good’ ones, not only from the citing paper but from the cited paper itself; this will especially be a problem for papers which have many citations to it.

Thirdly, we observe a general trend that the longer the citation context, the greater the retrieval effectiveness. Comparing the sentence-based contexts, `3sent` is ranked above `1sent` in 21 rankings, in 14 of which the differences were significant, compared to three rankings in which `1sent` is ranked higher, never significantly. The truncated versions of the sentence-based contexts are usually ranked lower: `1sentupto` is below `1sent` in 20 rankings, 10 times significantly, and above it in two rankings, neither of which are significant; `3sentupto` is below `3sent` in 21 rankings, 11 significant, and above it in three rankings, none significant. Thus, using neighbouring citations to delimit a citation’s context does not appear to be helpful; at least, not in the simplistic way we have tried here.

Finally, the window contexts also exhibit this trend: `win50` is usually ranked lowest of all the window contexts, being ranked beneath `win75` in 21 rankings (10 times significantly) and beneath `win100` in 22 rankings (nine times significantly), and is never significantly better than either `win75` or `win100`; `win75` is ranked lower than `win100` in 20 rankings, though the difference is only significant in two of these. In accordance with the trend, one of the longer window contexts, `win75` or `win100`, is usually ranked highest: in 19 out of 25 rankings. In five of the remaining six rankings, the difference between `win75` and `win100` and the top ranked context is insignificant. In the anomalous sixth case of Cosine bpref, only `full` is ranked significantly higher than either `win75` or `win100`. At a first glance, it appears that the increase in effectiveness with increasing window length is tailing off between `win75` and `win100`, as the improvement between these contexts is slightly smaller than between `win50` and `win75`. However, the difference between the average number of terms in `win75` and `win100` is smaller than that between `win50` and `win75` (42.0 versus 28.6) so the reduced improvement may simply be the result of the reduced increase in window size. Nevertheless, this trend of increasing effectiveness with increasing context length does not continue indefinitely, since effectiveness is decreased again by the time the entire citing paper is taken as the citation context: an optimal length of citation context exists somewhere between between `none` and `full`, though the contexts investigated here do not definitely show that optimum.

We now consider the relative effectiveness of the sentence-based and window contexts. Firstly, `win50` is usually ranked above `1sent`, in 19 of the 25 rankings, though the difference is usually insignificant (in 12 cases). Compared to `3sent` next, `win50` is ranked higher in slightly fewer (14) rankings and, in all 25 rankings, the difference between `win50` and `3sent` is insignificant. This initially seems to suggest that there is no advantage to making use of sentence boundaries for delimiting citation contexts: equivalent effectiveness can be achieved using a simple window method. However, the more effective `win50` is also longer on average than `3sent` (26.3 versus 98.4 terms) and, as we have seen, longer contexts tend to be more effective. Therefore, it is quite possible that sentence-based contexts are more effective than windows of equivalent length and that a longer sentence-based context would outperform `win50` and even the longer window contexts. Hence, as the effectiveness of increasingly longer window contexts tails off, the optimal context may, in fact, be a slightly shorter one constructed from sentences.

6.7 Comparison with other work

Our results show that indexing words from around citations to papers in combination with the words in the cited papers themselves improves retrieval effectiveness over indexing the paper alone. The experiments we have presented here are the first of their kind, in the way they make

use of the full text of citing and cited papers: previous experiments in this area have been limited in the extent to which they have used either the citing or cited papers, if not both.

Firstly, there have been experiments which have indexed cited papers using terms from citing papers but no terms from the cited papers themselves. Bradshaw's (2003) experiment is similar to ours in the way that it indexes cited papers using their citation contexts; the Citeseer window of one hundred words is like our `win50` context. However, Bradshaw indexes the citation contexts alone and compares this to indexing the cited paper alone; the combination of citation contexts with the cited paper is not explored. Similarly, Dunlop & van Rijsbergen (1993) create document representations of cited documents from their citing documents and compare retrieval using this representation to using the cited documents alone. Again, the combination of information from citing and cited documents is not explored. Furthermore, in Dunlop & van Rijsbergen's experiment, the documents are paper abstracts and not full papers; this is a somewhat outdated experiment, now that technology has advanced enough to allow full-text indexing of documents to become standard practice. Another difference is that Dunlop & van Rijsbergen index the whole of the citing abstracts, rather than the citation contexts specifically, as we do.

Secondly, O'Connor (1982) did investigate the combination of terms from citing and cited papers for retrieval. This experiment is similar to ours in that it compares retrieval using citation contexts together with an existing document representation to retrieval using the original document representation. Like in Bradshaw's experiment and in our own, the words taken from the citing papers for indexing were citation contexts. However, like in Dunlop & van Rijsbergen's experiment, the base document representation was not the full paper; in this case, it was the paper title, abstract and human index terms. Furthermore, since the experiment was conducted before machine-readable documents were readily available, the 'automatic' procedures for identifying the citation contexts were manually simulated for the experiment.

Thus, while previous work has gone some way towards showing the usefulness of using terms from citing papers for indexing cited papers, our experiments are unique in exploring the combination of words from citing papers with the full text of the cited documents. Furthermore, our experiments have been conducted on a realistic collection of thousands of documents, using a large number of queries.

Chapter 7

Conclusions

We conclude by summarising the contributions of this thesis and proposing some directions for future research. First, let us reiterate the motivation behind our work. The premise is that, when citing a paper, authors describe some part or aspect of the paper and that, intuitively, these descriptions – these *citation contexts* – should contain good index terms for the paper. Some work has been done in this area but no previous experiments have used both citing and cited papers to the fullest extent: some experiments have investigated using words from citing documents *alone* to represent a cited document, and compared this to using the cited document alone; others have combined citation contexts with an existing representation of the cited document, but not the full text.

Our first major contribution, therefore, is that we have presented results from retrieval experiments using a novel document representation for scientific papers: the combination of citation contexts and the full text of the cited paper. This is akin to indexing the anchor text of linking web pages in addition to the linked page itself, which has become an established practice in Web IR. In this way, the cited document's description of itself (i.e., its content) is combined with external descriptions from citing documents. We hypothesised that combining these descriptions would give an enhanced document representation compared to using the cited document alone; that indexing citation contexts in addition to the full text of a cited paper would result in better retrieval effectiveness. Our experimental results confirm this intuition: adding citation terms to the full cited paper does increase retrieval effectiveness. From this, we can conclude that citation contexts do contribute something to the document representation that helps papers to be found. This 'something' may be new terms that are not found in the cited papers themselves but that citing authors use to describe them; it may be repetition of the important terms in the paper, boosting their frequency and visibility; quite possibly, it is both, as we found by looking at citations to our case study paper from Ritchie et al. (2006*b*).

Additionally, we have shown that weighting citation terms higher relative to document terms generally improves retrieval effectiveness further. This indicates that the citation terms are somehow *more* important than the document terms, for the purposes of retrieval. Does this mean that citing authors are better at describing a paper than the paper's own author? That they are better at describing what is important in the paper? That their external perspectives of the paper more closely reflect what someone searching for the paper would query for? Or is this simply a product of the high concentration of important terms in citation contexts, due to their concise, summary nature?

We have also experimented with varying the precise context from which citation terms are

taken and shown that this does have a significant effect on retrieval effectiveness. The optimal context seems to exist somewhere between the citing sentence and the full citing paper, though further investigation is needed to establish more precisely where. It remains to be seen whether the optimal citation context is determined by sentence boundaries or yet more detailed linguistic cues, or whether simpler methods, like a fixed window, are generally sufficient to capture the useful context around citations. Thus, further experimentation is required to more accurately determine how best to extract the useful words from around citations. This is one of the limitations of our experiments; our database design restricted the contexts that we eventually investigated.

Our work has other limitations. With unlimited resources and the wonderful benefits of hindsight, we would alter the following aspects of our experimental set-up.

Test collection Our test collection would have complete relevance judgements or, at least, more complete than it does have. We would use a much deeper pool and a greater number of contributing systems, including runs that use our citation method.

Document processing We would improve our document processing so that, ideally, all references and citations are correctly identified and matched. Specific improvements we would implement would be to correct errors in the reference list segmentation from PTX, to expand our method to deal with reference and citation styles that occur rarely in the ACL Anthology and to use some sort of feedback loop between the reference and citation processors to signal when errors have been made (e.g., citations and/or references that have not been matched) and try to resolve these cases. We would also try to locate and annotate more specific information in the references, particularly the publication name since we use this information to identify which citation contexts should be extracted to our database.

Citation database We would find a better way of identifying references to ACL Anthology papers, i.e., one that is more robust than using simple string matching against a library of publication names. Preferrably, this method would be dynamic, since the ACL Anthology is a growing archive and new publications (e.g., one-off workshops) are continually added; the index pages from the archive web site could be downloaded periodically and used to automatically update the list of publications. Likewise, we would dynamically add new records to the database from the index pages when the archive is updated with new documents. We would also improve our method for matching references with database records, in particular by trying to resolve the cases where the title USBC from the reference does not match that of any database record, perhaps using some alternative fuzzy matching between the title in the reference and the titles in the database. We would design the database to allow more flexible experimentation with different citation contexts.

Experiments We would use a more sensible implementation of term weighting, e.g., by extending the Lemur code to allow weighting by XML field, as already implemented in Indri. We would investigate a wider range of weights and identify the optimal relative weighting of citation terms to document terms in each retrieval model. In our context experiments, we would compare longer contexts and find the optimal length of both fixed window and sentence contexts, and compare their effectiveness.

Nevertheless, we have established that adding citation contexts to the full text of the cited paper increases retrieval effectiveness. However, we can only speculate as to precisely what

it is about these citation contexts that has this effect. As a first step for future research, we propose to try to answer some of these speculations, by examining in detail the retrieval rankings from our experiments. Query-averaged retrieval effectiveness scores tell us whether or not a technique 'works' but they cannot explain *why*. By looking at the terms introduced by our citation context method at the individual document level, and at the consequent changes in the document rankings, we should gain a firmer understanding of what the citation contexts contribute that helps relevant papers to be retrieved. This should guide us as to the best way to implement our method and, also, direct future experimentation.

This brings us to the second major contribution of this thesis. We have created a new test collection of scientific papers, particularly suitable for experiments with citation information. The test collection has the full text of many cited and citing papers; this opens the door for broader experimentation than previous collections have allowed, e.g., comparisons between citation methods using the full cited paper versus the abstract alone. The document collection is a $\sim$9800 document snapshot of the the ACL Anthology digital archive of computational linguistics (CL) research papers. The query set consists of 82 research questions from CL papers, with an average of 11.4 judged relevant documents per query. We have experimentally validated its utility as a coparative evaluation tool, by exhibiting the stability of its system ranking, which is very similar to the ranking that might be expected from the literature. We intend to make the test collection freely available and hope that it will be a useful resource for the IR community and further afield.

The full-text TREC Genomics collection and the INEX collection of IEEE articles became available after our test collection effort was already underway. In future work, we would like to test our method on these collections too, to compare its effectiveness on different domains. Citation practices vary between disciplines so, for instance, what makes the optimal context may be slightly different for different domains; the Genomics and INEX collections will allow such differences to be investigated.

We would also like to extend our comparison of citation contexts to include more linguistic methods. In particular, a comparison with O'Connor's (1982) methods for accurately identifying citation contexts would make a worthwhile addition to the comparisons we have presented here; it would be interesting to see whether O'Connor's methods can, after all, be implemented fully automatically, now that machine-readable papers are widely available and problems such as sentence boundary detection and review article identification are practically 'solved'. Furthermore, we would like to conduct a wider comparison with previous work; in particular, with Bradshaw's (2003) and Dunlop & van Rijsbergen's (1993) methods, i.e., constructing document representations from citation contexts alone and from the abstracts of citing documents alone, respectively.

Our method is an application of *citation content analysis* to information retrieval: we take the explicit, contentful words from citation contexts and index them as part of the cited document. Citation content analysis is part of a broader family called *citation context analysis*, that could also be put to use in IR. For instance, our proposed examination of the retrieval rankings from our experiments might show that only contexts from certain types of citation contribute useful index terms, while some might be better off excluded from the indexing or treated differently, e.g., given a different weight. Automatic citation classification might improve our method by allowing us to discriminate between how different sorts of citations are used.

References

Agrawal, R., Rajagopalan, S., Srikant, R. & Xu, Y. (2003), Mining newsgroups using networks arising from social behavior, *in* 'Proceedings of the International World Wide Web Conference (WWW)'.

Altomari, P. J. & Currier, P. A., eds (1996), *Proceedings of the TIPSTER text program - Phase II*, Morgan Kaufmann Publishing.

Aslam, J. A., Pavlu, V. & Rei, C. (2006), The Hedge algorithm for metasearch at TREC 15, *in* 'Proceedings of the Fifteenth Text REtrieval Conference (TREC 2006)', National Institute of Standards and Technology, p. 557.

Aslam, J. A., Pavlu, V. & Yilmaz, E. (2006), A statistical method for system evaluation using incomplete judgements, *in* 'Proceedings of the 29th Annual ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)', ACM Press.

Ayres, F., Nielsen, L., Ridley, M. & Torsun, I. (1996), 'USBC (universal standard bibliographic code): Its origin and evolution', *Journal of Librarianship and Information Science* **28**(2), 83–91.

Bennett, G., Scholer, F. & Uitdenbogerd, A. (2008), A comparative study of probabilistic and language models for information retrieval, *in* 'Nineteenth Australasian Database Conference (ADC)', Vol. 75 of *CRPIT*, ACS, pp. 65–74.

Bharat, K. & Mihaila, G. A. (2001), When experts agree: using non-affiliated experts to rank popular topics, *in* 'Proceedings of the International World Wide Web Conference (WWW)', pp. 597–602.

Bradshaw, S. (2003), Reference directed indexing: Redeeming relevance for subject search in citation indexes., *in* 'Proceedings of Research and Advanced Technology for Digital Libraries (ECDL)', pp. 499–510.

Brin, S. & Page, L. (1998), 'The anatomy of a large-scale hypertextual Web search engine', *Computer Networks and ISDN Systems* **30**, 107–117.

Briscoe, E. & Carroll, J. (2002), Robust accurate statistical annotation of general text, *in* 'Proceedings of the International Conference on Language Resources and Evaluation (LREC)', pp. 1499–1504.

Brooks, T. A. (1985), 'Private acts and public objects: An investigation of citer motives', *Journal of the American Society for Information Science* **36**(4), 223–229.

Brownson, H. L. (1960), 'Research in handling scientific information', *Science* **132**(3444), 1922–1931.

Buckley, C. (1985), Implementation of the SMART information retrieval system, Technical Report TR85-686, Cornell University, Ithaca, NY, USA.

Buckley, C. & Voorhees, E. M. (2004), Retrieval evaluation with incomplete information, *in* 'Proceedings of the 27th Annual ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)', ACM Press, pp. 25–32.

Carterette, B. (2007), Robust test collections for retrieval evaluation, *in* 'Proceedings of the 30th Annual ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)', ACM Press, pp. 55–62.

Carterette, B., Allan, J. & Sitaraman, R. (2006), Minimal test collections for retrieval evaluation, *in* 'Proceedings of the 29th Annual ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)', ACM Press, pp. 268–275.

Chakrabarti, S., Dom, B., Raghavan, P., Rajagopalan, S., Gibson, D. & Kleinberg, J. (1998), Automatic resource compilation by analyzing hyperlink structure and associated text, *in* 'Proceedings of the International World Wide Web Conference (WWW)', pp. 65–74.

Cleverdon, C. (1960), Report on the first stage of an investigation into the comparative efficiency of indexing systems, Technical report, ASLIB Cranfield Project.

Cleverdon, C. (1997), The Cranfield tests on index language devices, *in* 'Readings in information retrieval', Morgan Kaufmann Publishers Inc., pp. 47–59.

Cleverdon, C., Mills, J. & Keen, M. (1966), Factors determining the performance of indexing sytems, volume 1. design, Technical report, ASLIB Cranfield Project.

Cooper, W. S. (1972), The inadequacy of probability of usefulness as a ranking criterion for retrieval system output. Unpublished working paper.

Councill, I. G., Giles, C. L. & Kan, M.-Y. (2008), ParsCit: An open-source CRF reference string parsing package, *in* 'Proceedings of the International Conference on Language Resources and Evaluation (LREC) (to appear)'.

Craswell, N., Hawking, D. & Robertson, S. (2001), Effective site finding using link anchor information, *in* 'Proceedings of the 24th Annual ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)', ACM Press, pp. 250–257.

Dunlop, M. D. & van Rijsbergen, C. J. (1993), 'Hypermedia and free text retrieval', *Information Processing and Management* **29**(3), 287–298.

Fairthorne, R. A. (1956), 'The patterns of retrieval', *American Documentation* **7**(2), 65–70.

Fujii, A. (2007), Enhancing patent retrieval by citation analysis, *in* 'Proceedings of the 30th Annual ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)', ACM Press, pp. 793–794.

Garcia, S., Lester, N., Scholer, F. & Shokouhi, M. (2006), RMIT university at TREC 2006: Terabyte track, *in* 'Proceedings of the Fifteenth Text REtrieval Conference (TREC 2006)', p. 623.

Garfield, E. (1972), 'Citation analysis as a tool in journal evaluation', *Science* **178**(4060), 471–479.

Garfield, E. (1979), *Citation indexing: its theory and application in science, technology, and humanities*, Wiley.

Garfield, E. (1997), 'Validation of citation analysis', *Journal of the American Society for Information Science* **48**(10), 962–964. Letter to the editor.

Gee, F. R., ed. (1999), *The proceedings of the TIPSTER text program - Phase III*, Morgan Kaufmann Publishing.

Gövert, N. & Kazai, G. (2002), Overview of the INitiative for the Evaluation of XML retrieval (INEX) 2002, *in* 'Proceedings of the Workshop of the INitiative for the Evaluation of XML Retrieval (INEX)', pp. 1–17.

Harman, D. K. (1992), Relevance feedback revisited, *in* 'Proceedings of the 15th Annual ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)', ACM Press, pp. 1–10.

Harman, D. K. (2005), The TREC test collections, *in* E. M. Voorhees & D. K. Harman, eds, 'TREC Experiment and Evaluation in Information Retrieval', MIT Press, chapter 2.

Hawking, D. & Craswell, N. (2005), The very large collection and web tracks, *in* E. M. Voorhees & D. K. Harman, eds, 'TREC: Experiment and Evaluation in Information Retrieval', MIT Press, chapter 9.

Hersh, W. & Bhupatiraju, R. T. (2003), TREC genomics track overview, *in* 'Proceedings of the Twelfth Text REtrieval Conference (TREC 2003)', National Institute of Standards and Technology, pp. 14–23.

Hersh, W., Cohen, A. M., Roberts, P. & Rekapilli, H. K. (2006), TREC 2006 genomics track overview, *in* 'Proceedings of the Fifteenth Text REtrieval Conference (TREC 2006)', National Institute of Standards and Technology.

Holsti, O. R. (1969), *Content Analysis for the Social Sciences and Humanities*, Addison Wesley, Reading, MA.

Hotho, A., Jäschke, R., Schmitz, C. & Stumme, G. (2006), Information retrieval in folksonomies: Search and ranking, *in* 'The Semantic Web: Research and Applications', Vol. 4011/2006 of *Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, pp. 411–426.

Järvelin, K. & Kekäläinen, J. (2002), 'Cumulated gain-based evaluation of ir techniques', *ACM Transactions on Information Systems* **20**(4), 422–446.

Jones, S., Cunningham, S. J. & McNab, R. (1998), Usage analysis of a digital library, *in* 'Proceedings of the third ACM conference on Digital Libraries', ACM Press, pp. 293–294.

Joseph, M. T. & Radev, D. R. (2007), Citation analysis, centrality, and the ACL Anthology, Technical Report CSE-TR-535-07, University of Michigan. Department of Electrical Engineering and Computer Science.

Katter, R. V. (1968), 'The influence of scale form on relevance judgments', *Information Storage and Retrieval* **4**(1), 1–11.

Kessler, M. M. (1963), 'Bibliographic coupling between scientific papers', *American Documentation* **14**(1), 10–25.

Kleinberg, J. M. (1999), 'Authoritative sources in a hyperlinked environment', *Journal of the ACM* **46**(5), 604–632.

Kluck, M. (2003), The GIRT data in the evaluation of CLIR systems - from 1997 until 2003, *in* 'Proceedings of the Cross-Language Evaluation Forum Workshop (CLEF)', pp. 376–390.

Krippendorff, K. (2004), *Content Analysis: An Introduction to Its Methodology*, 2 edn, Sage Publications Inc.

Krovetz, R. (1993), Viewing morphology as an inference process,, *in* 'Proceedings of the 16th Annual ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)', ACM Press, pp. 191–203.

Lawrence, S., Bollacker, K. & Giles, C. L. (1999), Indexing and retrieval of scientific literature, *in* 'Proceedings of the 8th Conference on Information and Knowledge Management (CIKM)', ACM Press, pp. 139–146.

Lewin, I., Hollingsworth, B. & Tidhar, D. (2005), Retrieving hierarchical text structure from typeset scientific articles - a prerequisite for e-science text mining, *in* 'Proceedings of the UK e-Science All Hands Meeting'.

Liu, M. (1993), 'The complexities of citation practice: a review of citation studies', *Journal of Documentation* **49**(4), 370–408.

Malik, S., Trotman, A., Lalmas, M. & Fuhr, N. (2006), Overview of INEX 2006, *in* 'Proceedings of the Workshop of the INitiative for the Evaluation of XML Retrieval (INEX)'.

Marchiori, M. (1997), 'The quest for correct information on the Web: Hyper search engines', *Computer Networks and ISDN Systems* **29**(8–13), 1225–1236.

Maron, M. E. & Kuhns, J. L. (1960), 'On relevance, probabilistic indexing and information retrieval', *Journal of the ACM* **7**(3), 216–244.

McBryan, O. (1994), GENVL and WWWW: Tools for taming the web, *in* 'Proceedings of the International World Wide Web Conference (WWW)'.

Meij, E. & de Rijke, M. (2007), Using prior information derived from citations in literature search, *in* 'Proceedings of the International Conference on Recherche d'Information Assistée par Ordinateur (RIAO)'.

Merchant, R. H., ed. (1994), *Proceedings of the TIPSTER text program - Phase I*, Morgan Kaufmann Publishing.

Mizzaro, S. (1997), 'Relevance: The whole history', *Journal of the American Society for Information Science* **48**(9), 810–832.

Mizzaro, S. & Robertson, S. (2007), HITS hits TREC – exploring IR evaluation results with network analysis, *in* 'Proceedings of the 30th Annual ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)', ACM Press, pp. 479–486.

Mooers, C. (1951), Information retrieval viewed as temporal signalling, *in* 'Proceedings of the International Congress of Mathematicians', pp. 572–573.

Moravcsik, M. J. & Murugesan, P. (1975), 'Some results on the function and quality of citations', *Social Studies of Science* **5**, 88–91.

Nakov, P. I., Schwartz, A. S. & Hearst, M. A. (2004), Citances: Citation sentences for semantic analysis of bioscience text, *in* 'Proceedings of the SIGIR Workshop on Search and Discovery in Bioinformatics', ACM Press.

Nanba, H. & Okumura, M. (1999), Towards multi-paper summarization using reference information, *in* 'Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence', Morgan Kaufmann Publishers Inc., pp. 926–931.

Nanba, H. & Okumura, M. (2005), Automatic detection of survey articles, *in* 'Proceedings of Research and Advanced Technology for Digital Libraries (ECDL)', pp. 391–401.

Ntoulas, A., Cho, J. & Olston, C. (2004), What's new on the web? the evolution of the web from a search engine perspective, *in* 'Proceedings of the International World Wide Web Conference (WWW)'.

O'Connor, J. (1982), 'Citing statements: Computer recognition and use to improve retrieval', *Information Processing and Management* **18**(3), 125–131.

O'Connor, J. (1983), 'Biomedical citing statements: Computer recognition and use to aid full-text retrieval', *Information Processing and Management* **19**, 361–368.

Ogilvie, P. & Callan, J. (2001), Experiments using the Lemur toolkit, *in* 'Proceedings of the Tenth Text REtrieval Conference (TREC 2001)', National Institute of Standards and Technology, pp. 103–108.

Pitkow, J. & Pirolli, P. (1997), Life, death, and lawfulness on the electronic frontier, *in* 'Proceedings of the Conference on Human Factors in Computing Systems (CHI)'.

Powley, B. & Dale, R. (2007), Evidence-based information extraction for high accuracy citation and author name identification, *in* 'Proceedings of the International Conference on Recherche d'Information Assistée par Ordinateur (RIAO)'.

Ridley, M. J. (1992), 'An expert system for quality control and duplicate detection in bibliographic databases', *Program: Automated Library and Information Systems* **26**(1), 1–18.

Ritchie, A., Robertson, S. & Teufel, S. (2007), Creating a test collection: Relevance judgements of cited & non-cited papers, *in* 'Proceedings of the International Conference on Recherche d'Information Assistée par Ordinateur (RIAO)'.

Ritchie, A., Teufel, S. & Robertson, S. (2006*a*), Creating a test collection for citation-based IR experiments, *in* 'Proceedings of the Human Language Technology conference and North American chapter of the Association for Computational Linguistics annual meeting (HLT-NAACL)', pp. 391–398.

Ritchie, A., Teufel, S. & Robertson, S. (2006*b*), How to find better index terms through citations, *in* 'Proceedings of the COLING/ACL Workshop How Can Computational Linguistics Improve Information Retrieval?'.

Robertson, S. (2007), On document populations and measures of IR effectiveness, *in* S. Dominich & F. Kiss, eds, 'Proceedings of the 1st International Conference on the Theory of Information Retrieval (ICTIR'07)', Foundation for Information Society, pp. 9–22.

Robertson, S. E. (1977), 'The probability ranking principle in IR', *Journal of Documentation* **33**, 294–304.

Robertson, S., Zaragoza, H. & Taylor, M. (2004), Simple BM25 extension to multiple weighted fields, *in* 'Proceedings of the 13th Conference on Information and Knowledge Management (CIKM)', ACM Press, pp. 42–49.

Salton, G. (1992), 'The state of retrieval system evaluation', *Information Processing and Management* **28**(4), 441–449.

Salton, G., Wong, A. & Yang, C. (1975), 'A vector space model for automatic indexing', *Communications of the ACM* **18**(11), 613–620.

Sanderson, M. & Zobel, J. (2005), Information retrieval system evaluation: effort, sensitivity, and reliability, *in* G. Marchionini, A. Moffat, J. Tait, R. Baeza-Yates & N. Ziviani, eds, 'Proceedings of the 28th Annual ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)', ACM Press, pp. 162–169.

Saracevic, T. (1975), '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.

Saracevic, T., Kantor, P. B., Chamis, A. Y. & Trivison, D. (1988), 'A study of information seeking and retrieving. Parts I–III', *Journal of the American Society for Information Science* **39**(3), 161–216.

Schamber, L., Eisenberg, M. B. & Nilan, M. S. (1990), 'A re-examination of relevance: Toward a dynamic, situational definition', *Information Processing and Management* **26**, 755–776.

Schneider, J. (2004), Verification of bibliometric methods' applicability for thesaurus construction, PhD thesis, Department of Information Studies, Royal School of Library and Information Science.

Small, H. (1973), 'Co-citation in the scientific literature: A new measurement of the relationship between two documents', *Journal of the American Society of Information Science* **24**(4), 265–269.

Small, H. (1982), Citation context analysis, *in* B. Dervin & M. J. Voigt, eds, 'Progress in Communication Sciences', Vol. 3, Ablex Publishing, pp. 287–310.

Small, H. G. (1978), 'Cited documents as concept symbols', *Social Studies of Science* **8**(3), 327–340.

Spärck Jones, K. (1981), The Cranfield tests, *in* K. Spärck Jones, ed., 'Information Retrieval Experiment', Butterworths, chapter 13, pp. 256–284.

Spärck Jones, K. (1990), What sort of thing is an AI experiment?, *in* D. Partridge & Y. Wilks, eds, 'The Foundations of Artificial Intelligence: A Sourcebook', Cambridge University Press, Cambridge, UK.

Spärck Jones, K. & van Rijsbergen, C. J. (1976), 'Information retrieval test collections', *Journal of Documentation* **32**(1), 59–75.

Spiegel-Rösing, I. (1977), 'Science studies: Bibliometric and content analysis', *Social Studies of Science* **7**(1), 97–113.

Stirling, K. H. (1975), The Effect of Document Ranking on Retrieval System Performance: A Search for an Optimal Ranking Rule, PhD thesis, University of California, Berkeley.

Strohman, T., Croft, W. B. & Jensen, D. (2007), Recommending citations for academic papers, *in* 'Proceedings of the 30th Annual ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)', ACM Press, pp. 705–706.

Strohman, T., Metzler, D., Turtle, H. & Croft, W. B. (2005), Indri: a language-model based search engine for complex queries, Technical report, Center for Intelligent Information Retrieval, University of Massachusetts.

Swales, J. (1986), 'Citation analysis and discourse analysis', *Applied Linguistics* **7**(1), 39–56.

Taghva, K., Coombs, J., Pareda, R. & Nartker, T. (2004), Language model-based retrieval for Farsi documents, *in* 'Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC)', Vol. 2, pp. 13–17.

Tague-Sutcliffe, J. (1992), 'The pragmatics of information retrieval experimentation, revisited', *Information Processing and Management* **28**(4), 467–490.

Taylor, R. (1968), 'Question negotiation and information seeking in libraries', *College and Research Libraries* **29**(3), 178–194.

Teevan, J., Alvarado, C., Ackerman, M. S. & Karger, D. R. (2004), The perfect search engine is not enough: a study of orienteering behavior in directed search, *in* 'Proceedings of the Conference on Human Factors in Computing Systems (CHI)', ACM Press, pp. 415–422.

Teufel, S. (1999), Argumentative Zoning: Information Extraction from Scientific Text, PhD thesis, School of Cognitive Science, University of Edinburgh, UK.

Teufel, S. & Elhadad, N. (2002), Collection and linguistic processing of a large-scale corpus of medical articles, *in* 'Proceedings of the International Conference on Language Resources and Evaluation (LREC)'.

Teufel, S., Siddharthan, A. & Tidhar, D. (2006), Automatic classification of citation function, *in* 'Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)', pp. 103–110.

Thomas, P. & Hawking, D. (2006), Evaluation by comparing result sets in context, *in* 'Proceedings of the 15th Conference on Information and Knowledge Management (CIKM)', ACM Press, pp. 94–101.

Vickery, B. C. (1967), 'Reviews of CLEVERDON, C. W., MILLS, J. and KEEN, E. M. the Cranfield 2 report', *Journal of Documentation* **22**, 247–249.

Voorhees, E. M. (1998), Variations in relevance judgments and the measurement of retrieval effectiveness, *in* 'Proceedings of the 21st Annual ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)', ACM Press, pp. 315–323.

Voorhees, E. M. (2001), Evaluation by highly relevant documents, *in* 'Proceedings of the 24th Annual ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)', ACM Press, pp. 74–82.

Voorhees, E. M. (2002), The philosophy of information retrieval evaluation, *in* 'Proceedings of the Cross-Language Evaluation Forum Workshop (CLEF)', pp. 355–370.

Voorhees, E. M. (2005), Overview of the TREC 2005 robust retrieval track, *in* 'Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005)', National Institute of Standards and Technology.

Voorhees, E. M. & Harman, D. (1999), Overview of the eighth Text REtrieval Conference (TREC-8), *in* 'Proceedings of the Eighth Text REtrieval Conference (TREC-8)'.

Voorhees, E. M. & Harman, D. K., eds (2005), *TREC Experiment and Evaluation in Information Retrieval*, MIT Press.

Walker, D. J., Clements, D. E., Darwin, M. & Amtrup, J. W. (2001), Sentence boundary detection: A comparison of paradigms for improving MT quality, *in* 'Proceedings of the Machine Translation Summit VIII'.

Wasserman, S. & Faust, K. (1995), *Social Network Analysis: Methods and Applications*, Structural Analysis in the Social Sciences, Cambridge University Press.

White, H. D. (2004), 'Citation analysis and discourse analysis revisited', *Applied Linguistics* **1**, 89–116.

Yu, B. & Singh, M. P. (2000), A social mechanism of reputation management in electronic communities, *in* 'Proceedings of the 4th International Workshop on Cooperative Information Agents', Springer-Verlag, pp. 154–165.

Zaragoza, H., Craswell, N., Taylor, M., Saria, S. & Robertson, S. (2004), Microsoft Cambridge at TREC-13: Web and HARD tracks, *in* 'Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004)', National Institute of Standards and Technology.

Zhai, C. X., Cohen, W. W. & Lafferty, J. (2003), Beyond independent relevance: methods and evaluation metrics for subtopic retrieval, *in* 'Proceedings of the 26th Annual ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)', ACM Press, pp. 10–17.

Zhou, Y. & Croft, W. B. (2007), Query performance prediction in web search environments, *in* 'Proceedings of the 30th Annual ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)', ACM Press, pp. 543–550.

Zobel, J. (1998), How reliable are the results of large-scale information retrieval experiments?, *in* 'Proceedings of the 21st Annual ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)', ACM Press, pp. 307–314.

Appendix A

Feasibility study

A.1 Ideal citation term ranking by TF*IDF

Rank			
Ideal	**Doc**	**TF*IDF**	**Term**
1	1	351.73	french
2	2	246.52	alignments
3	3	238.39	fertility
4	4	212.20	alignment
5	5	203.28	cept
6	8	158.45	probabilities
7	9	150.74	translation
8	12	106.11	model
9	17	79.47	probability
10	18	78.37	models
11	19	78.02	english
12	21	76.23	parameters
13	24	71.77	connected
14	28	62.48	words
15	32	57.57	em
13	35	54.88	iterations
14	45	45.00	statistical
15	54	38.25	training
16	69	32.93	word
17	74	31.31	pairs
18	81	29.29	machine
19	83	28.53	empty
20	130	19.72	series

A.2 Term ranking changes (ideal and fixed window)

Term	TF*IDF Δ	TF*IDF Doc+ideal	Ideal rank Δ
ibm	24.24	37.46	28 → 20
generative	4.44	11.10	38 → 33
source	5.35	6.42	65 → 44
decoders	6.41	6.41	__ → 45
corruption	6.02	6.02	__ → 46
expectation	2.97	5.94	51 → 47
relationship	2.96	5.92	52 → 48
story	2.94	5.88	53 → 49
noisy-channel	5.75	5.75	__ →52
extract	1.51	7.54	41 → 38

Term	TF*IDF Δ	TF*IDF Doc+fixed	Ideal rank Δ
ibm	48.48	61.70	28 → 18
target	19.64	19.64	__→ 26
source	14.99	16.06	65 → 32
phrase-based	14.77	14.77	__→ 36
trained	14.64	19.52	43 → 27
approaches	11.03	11.03	__ → 41
parallel	9.72	17.81	34 → 29
generative	8.88	15.54	38 → 33
train	8.21	8.21	__ → 45
channel	6.94	6.94	__ → 55
expectation	5.93	8.90	51 → 44
learn	5.93	7.77	60 → 47

A.3 New non-zero TF*IDF terms (ideal and fixed window)

Term	TF*IDF
decoders	6.41
corruption	6.02
noisy-channel	5.75
attainable	5.45
target	5.24
source-language	4.99
phrase-based	4.92
target-language	4.82
application-specific	4.40
train	4.10
intermediate	4.01
channel	3.47
approaches	3.01
combinations	1.70
style	2.12
add	1.32
major	1.16
due	0.83
considered	0.81
developed	0.78

Term	TF*IDF
target	19.64
phrase-based	14.77
approaches	11.03
train	8.21
channel	6.94
decoders	6.41
corruption	6.02
noisy-channel	5.75
attainable	5.45
source-language	4.99
target-language	4.82
application-specific	4.40
intermediate	4.01
combinations	3.40
style	2.12
considered	1.62
major	1.16
due	0.83
developed	0.78

A.4 ‘Noisy’ fixed window terms

(Overleaf.)

TF	# terms	Terms
13	1	asr
8	4	caption, closed, section, methods
7	2	method, sentences
6	4	describes, example, languages, system
5	6	corpus, dictionary, heuristic, large, paper, results
4	17	account, aligned, confidence, dependency, details, during, equation, generally, given, manual, measures, order, probabilistic, proposed, shown, simplified, systems, word-aligned
3	29	according, algorithm, applications, build, case, choosing, chunk, current, described, employed, equivalence, experiments, introduced, introduction, length, links, number, obtain, obtained, performance, performing, problem, produced, related, show, sum, true, types, work
2	64	adaptation, akin, approximate, bitext, calculated, called, categories, certain, chunks, common, consider, consists, domain-specific, error, estimation, experimental, extracted, families, feature, features, found, functions, generated, generic, giza, good, high, improve, information, input, iraq, knowledge, large-scale, lexicon, linked, log-linear, maximum, measure, notion, omitted, original, output, parameter, pick, position, practice, presents, quality, rate, represented, researchers, rock, role, sinhalese, takes, tamil, text-to-text, toolkit, transcripts, transcriptions, translations, version, word-based, word-to-word
1	252	access, accuracy, achieve, achieving, actual, addition, address, adopted, advance, advantages, aligning, amalgam, annotated, applied, apply, applying, approximated, association, asymmetric, augmented, availability, available, average, back-off, base, baum-welch, begin, bitexts, bunetsu, candidate, candidates, cat, central, chinese, choose, chunk-based, class, closely, collecting, combination, compare, compared, compares, computed, concludes, consequently, contributed, convention, corpora, correspondence, corrupts, cost, counts, coverage, crucial, currently, decades, decoding, defines, denote, dependent, depending, determine, dictionaries, direct, directions, disadvantages, distinction, dominated, dynamic, efforts, english-chinese, english-spanish, enumerate, eojeol, eq, equations, errors, evaluation, excellent, expansion, explicitly, extracts, failed, fairly, final, finally, fit, flat-start, followed, form, formalisms, formulation, generation, gis, give, grouped, hallucination, halogen, handle, heuristic-based, hidden, highly, hill-climbing, hmm-based, hypothesis, ideal, identified, identify, identity, immediate, implemented, improved, improves, incorporate, increase, influence, initial, initialize, inspired, interchanging, introduces, investigations, involve, kate, kind, learning, learns, letter, letters, lexical, likelihood, link, list, longer, lowercase, main, make, makes, mapping, maximal, maximizes, means, modeling, modified, names, needed, nitrogen, nodes, occupy, omitting, optimal, outperform, overcome, parse, parser, part, part-of-speech, path, performed, play, plays, popular, pos, positions, power, precision, probable, produce, programming, promising, real-valued, reason, recall, recent, recently, recognition, recursion, recursively, reduction, reductions, refine, relative, relying, renormalization, representation, require, requires, research, restricting, reveal, sample, sampling, satisfactory, segments, semantic, sequences, setting, shortcomings, showed, significant, significantly, similarity, similarly, simple, simplicity, situation, space, speech, spelling, state-of-the-art, step, strategies, string, strong, studies, summaries, summarization, supervised, syntactic, tags, task-specific, technique, techniques, technologies, terms, testing, threshold, translation-related, transliteration, tree, trees, trellis, type, underlying, unrealistic, unsupervised, uppercase, value, viterbi, wanted, ways, well-formedness, well-founded, widely, widespread, works, written, wtop, yasmet, years, yields

Appendix B

Test collection

B.1 Methodology comparison with Cranfield 2

Feature	Cranfield 2	ACL Anthology
Document collection	Manufactured.	Existing.
	High speed aerodynamics and aircraft structures papers.	Computational linguistics papers.
	1400 papers.	∼10,000 papers.
	All in English.	Vast majority in English, non-English papers removed.
Query source papers	Mostly published within 1.5 years (1962-3).	All published in same year (2005).
	Mostly articles from one prominent journal, some research reports.	All papers presented at two main conferences.
	Mostly American publications (76.9%), some British (22.5%), few Swedish (0.6%).	All from international conferences → international authors and institutions.
	All in English (c.f. base document selection criteria).	All in English.
Phase One methodology	Asked for relevance judgements for up to ten references.	Asked for relevance judgements for all references.
	Asked for 'no more than three supplementary questions'.	Asked for any number of additional research questions, with no limit imposed, only suggested (indirectly, by giving space for three on form).

Phase One returns	Sent 271 letters plus 'chase' letters later for those not replied.	Sent over 300 invitations plus reminder emails for those not replied.
	182 completed forms (67.2%)	89 completed forms (28.3%)
	641 questions (3.5 per author)	258 questions (2.9 per author), one with no judgements
Query filtering	Selected questions with two or more relevant (grades 1,2,3) references and 'grammatically complete'.	Kept questions with two or more relevant (grades 2,3,4) references, one or more relevant Anthology reference, with no restrictions on grammaticality. Discarded questions from co-authors whose first author had also replied.
	360 questions remaining.	198 questions remaining.
Query reformulations	Resolved anaphoric phrases ('inserted missing words').	Resolved anaphoric phrases (using words from previous questions) and added context words from previous questions where context seemed to be assumed.
	?	Fixed typographical errors.
	?	Removed contentless 'filler' phrases or repetitions of existing content.
	?	35 questions reformulated.
	Resubmitted to authors in second round of judgements. No disagreement with the amendments.	Asked for approval in Phase Two. All approved except two, for which new reformulations were given by the authors. These were resubmitted to the pooling process.
Manual searches	Performed by five postgraduate students with knowledge of aerodynamics.	Performed by one postgraduate student with knowledge of computational linguistics.
	1500 person-hours.	$\sim$80 person-hours ($>$180 queries, aimed for 15 minutes each but usually spent longer).
	Examined every document in collection.	Google searched document collection via Anthology website.
	Had access to author response forms, relevance judgements, source paper, bibliographic details of the document collection and the complete documents.	Had access to equivalent materials.
	Instructed to make 'liberal' judgements.	Liberality of relevance judgements depended on specificity of query and how many relevant documents there seemed to be.
	No claims made that every possible relevant document was found.	Agreed! (Reassuring to find overlaps between manual search results and both judged relevant references and automatic search results.)
	3.3 possibly relevant documents found per query.	10.7 possibly relevant documents found per query.

Phase Two methodology	Used previous relevance scale.	Changed to binary relevance.
	Asked authors to grade the relative importance of each search 'term or concept', to list alternative terms/concepts and, if necessary, to include a completely rephrased version.	(Specific to index language tests so outwith our interest)
	Sent authors photocopy of original form, new document list for judgement (authors, titles, references, which question thought relevant to, abstracts). 11 sheets sent on average.	Sent one email, comprising invitation letter, instructions, one attachment (PDF form and instructions), one form with link to webpage per query.
Phase Two returns	144 out of 182 authors returned completed forms (79.1%)	44 out of 74 completed forms were returned (59.5%)
	Received judgements for 201 out of 283 queries (71.0%). 78 queries with no possibly relevant documents found (so not resubmitted to authors). 279 queries in total (not 221?).	Received judgements for 83 out of 183 queries (45.4%). 82 queries with Phase One and Two judgements.
	Detailed analysis of judgements coming from different methods of finding documents e.g. references vs manual vs automatic, how many of which relevance grades etc.	Analysis reported in Ritchie, Robertson & Teufel (2007).
Automatic methods	Bibliographic coupling: include all documents with seven or more references in common with one of author's cited relevant (grade 1,2,3) papers.	Pooling: include top documents retrieved by three standard IR models (Okapi BM25, KL-divergence, Cosine similarity).
	Small overlap between manual and automatic results (15 out of 213).	Some overlap between manual and automatic results. See Ritchie et al. (2007).
	Submit all possibly relevant documents to authors for relevance judgement.	Submit top 15 possibly relevant documents to authors for relevance judgement. All manual search results were included (even if there were more than 15) then one from each of the automatic lists (removing duplicates) up to 15.

B.2 Conference author materials

B.2.1 Author invitation (Phase One)

Dear Dr Choi,

I would like to invite you, as the author of a recent computational linguistics paper, to participate in an endeavour to work towards a collection of search queries and relevance assessments. Our chosen corpus is the ACL Anthology.

I am a PhD student at Cambridge University Computer Laboratory, where my research interests focus on scientific citations in information retrieval (IR), working with articles from the ACL Anthology. In order to evaluate IR techniques, however, a set of search queries is needed, as well as assessments as to which documents in the corpus are relevant to those queries.

We assume that every paper has an underlying research question it aimed to answer and that constitutes a valid search query. Not only is noone better qualified to put into words those original research questions than the paper authors themselves, but neither is anyone better qualified to decide what other papers are relevant to that question: you are the experts in the field.

I would be extremely grateful if you would agree to partipate. The hope is that, if enough authors can spare twenty minutes or so to send us the information asked for, we can create a usable, useful addition to the ACL Anthology. Please find attached a) instructions on how to participate and b) a personalized response form. Alternatively, I would be happy to send you a paper copy, if you would prefer. Your contribution can be emailed back to me (using the ASCII form at the bottom of this email) or paper copies of the form posted or faxed to me at

Anna Ritchie
University of Cambridge
Computer Laboratory
William Gates Building
15 JJ Thomson Avenue
Cambridge CB3 0FD
UK
Fax: +44 (0)1223 334678

There will be a follow up stage to this data collection but participation in this stage is optional; any help you can give now will be valuable independently of later participation.

Many thanks for your time and congratulations on the acceptance of your paper to ACL 2005,

Anna Ritchie

****

INSTRUCTIONS FOR AUTHORS:
FORMULATING QUESTIONS AND MAKING RELEVANCE ASSESSMENTS

The purpose of this experiment is to build a collection of queries and relevance assessments for a corpus of computational linguistics papers. You will be asked to write down the problem(s) that your paper dealt with and to assess how relevant each of the references in your reference list are in relation to those problems.

1) Main Problem

Write down, in the form of a question (one sentence), the basic problem your paper addressed, i.e., that was the focus of your work.

2) Further Problems

Write down any additional/subsidiary/subordinate problems that affected your work, if any. These might be more general problems, the solution of which your research contributed to, or subproblems...

* for which it was necessary to find new or existing solutions to in order to carry out the main research.
* relating to the methodology used in the work, rather than the theoretical research problem.

(It is perfectly possible that your paper was influenced by only one problem, in which case, proceed to step 3.)

3) Reference List Relevance Assessments

In tabular form, for each reference in your paper's reference list, assign a score to indicate the relevance of that reference to each of the problems you have written down for 1 and 2, using the relevance scale given below.

****

AUTHOR RESPONSE FORM

Yejin Choi

''Identifying Sources of Opinions with Conditional Random Fields and Extraction Patterns''

1) Main Problem

2) Further Problems

i
ii
iii

3) Relevance Assessments

REFERENCES	PROBLEM			
	Main	i	ii	iii
C. Baker et al, 1998				
S. Bethard et al, 2004				
D. Bikel et al, 1997				
E. Breck and C. Cardie, 2004				
C. Cardie et al, 2004				
M. Collins, 1999				
H. Cunningham et al, 2002				
S. Das and M. Chen, 2001				
K. Dave et al, 2003				
J. Lafferty et al, 2001				
B. Levin, 1993				
A. K. McCallum, 2002				
A. K. McCallum, 2003				
A. K. McCallum and W. Li, 2003				
S. Morinaga et al, 2005				
M. Palmer et al, 2005				
B. Pang et al, 2002				
B. Pang and L. Lee, 2004				
L. A. Ramshaw and M. P. Marcus, 1995				
E. Riloff, 1996a				
E. Riloff, 1996b				
E. Riloff and J. Wiebe, 2003				
E. Riloff and W. Phillips, 2004				
S. Sarawagi and W. W. Cohen, 2004				
P. Turney, 2002				
T. Wilson et al, 2004				
T. Wilson et al, 2005				
J. Yi et al, 2003				

H. Yu and V. Hatzivassiloglou, 2003
J. Wiebe et al, 2002
J. Wiebe and E. Riloff, 2005
J. Wiebe et al, 2005

****

RELEVANCE SCALE

Relevance 4
The reference is crucially relevant to the problem. Knowledge of the contents of the referred work will be fundamental to the reader's understanding of your paper. Often, such relevant references are afforded a substantial amount of text in a paper e.g., a thorough summary.

* In the case of subproblems, the reference may provide a complete solution (e.g., a reference explaining an important tool used or method adopted for the research)
* In the case of the main problem, the reference may provide a complete solution (e.g., an existing, alternative solution to the problem that your work directly contrasts or proves incorrect).
* In either case, the reference may provide a partial solution that your work builds upon (e.g., previous work of your own or others that your current work extends or improves).

Relevance 3
The reference is relevant to the problem. It may be helpful for the reader to know the contents of the referred work, but not crucial. The reference could not have been substituted or dropped without making significant additions to the text. A few sentences may be associated with the reference.

* The reference may be the standard reference given for a particular tool or method used, of which an understanding is not necessarily required to follow your paper.
* The referred work may give an alternative approach to the problem that is not being directly compared in the current work.
* The referred work may give an approach to a similar or related problem.

Relevance 2 2
The reference is somewhat (perhaps indirectly) relevant to the problem. Following up the reference probably would not improve the reader's understanding of your paper. Alternative references may have been equally appropriate (e.g., the reference was chosen as a representative example from a number of similar references or included in a list of similar references). Or the reference could have been dropped without damaging the informativeness of your paper. Minimal text will be associated with the reference.

* The reference may be included to give some historical background to the problem.
* The reference may be included to acknowledge a (non-critical) contribution.

Relevance 1
The reference is irrelevant to this particular problem.

* E.g., a reference about an implementation strategy may be irrelevant to a subproblem about evaluation strategy.

B.2.2 Author response form (Phase One)

AUTHOR RESPONSE FORM

Yejin Choi

''Identifying Sources of Opinions with Conditional Random Fields and Extraction

Patterns''

1) Main Problem

How can we handle the problem of automatic identification of sources of opinions?

2) Further Problems

i how can we model the problem with conditional random fields?
ii how can we model the problem with extraction patterns?
iii how can we train and test our model?

3) Relevance Assessments

REFERENCES	PROBLEM			
	Main	i	ii	iii
C. Baker et al, 1998	2	1	1	1
S. Bethard et al, 2004	3	1	1	2
D. Bikel et al, 1997	2	2	1	1
E. Breck and C. Cardie, 2004	2	1	1	1
C. Cardie et al, 2004	2	1	1	2
M. Collins, 1999	1	2	1	1
H. Cunningham et al, 2002	1	2	1	1
S. Das and M. Chen, 2001	2	1	1	1
K. Dave et al, 2003	2	1	1	1
J. Lafferty et al, 2001	3	4	1	2
B. Levin, 1993	2	3	1	2
A. K. McCallum, 2002	1	4	1	2
A. K. McCallum, 2003	1	3	1	2
A. K. McCallum and W. Li, 2003	1	2	1	1
S. Morinaga et al, 2005	2	1	1	1
M. Palmer et al, 2005	2	1	1	1
B. Pang et al, 2002	2	1	1	1
B. Pang and L. Lee, 2004	2	1	1	1
L. A. Ramshaw and M. P. Marcus, 1995	1	2	1	1
E. Riloff, 1996a	3	1	4	2
E. Riloff, 1996b	3	1	4	2
E. Riloff and J. Wiebe, 2003	2	1	1	1
E. Riloff and W. Phillips, 2004	3	1	4	2
S. Sarawagi and W. W. Cohen, 2004	2	1	1	1
P. Turney, 2002	2	1	1	1
T. Wilson et al, 2004	2	1	1	1
T. Wilson et al, 2005	2	1	1	1
J. Yi et al, 2003	2	1	1	1
H. Yu and V. Hatzivassiloglou, 2003	2	1	1	1
J. Wiebe et al, 2002	3	2	2	3
J. Wiebe and E. Riloff, 2005	2	1	1	1
J. Wiebe et al, 2005	3	2	2	3

B.2.3 Author invitation (Phase Two)

Dear Dr Choi,

Once again, thank you for the information you contributed to our test collection for the ACL Anthology. Your data has been processed, laying the foundations for what we hope will be a useful, high quality resource for the research community. We are now entering the second and final stage of the collection, where we attempt to expand on the relevance information you have already provided, by asking you to judge the relevance of papers outside your reference list with respect to your research question(s). To save your time, we ask only for binary relevance judgements this time: relevant or irrelevant. I would be extremely grateful if you could spare some time to make these additional judgements; they will greatly increase the collection's quality.

Below you will find a response form with, for each of the research questions you submitted, a list of papers for which we request your relevance judgements. These papers are the result of pooling the

outputs from a variety of standard retrieval algorithms, as well as manual searching, using your research question as a search query. The papers are available for your inspection at a personalized web page (c.f. below). Instructions on how to make the relevance judgements are given below and in the PDF attachment.

If you find you do not have time to make all of these judgements, judgements for just some of your questions would still be very useful to us. Your questions are presented in priority order.

We have reformulated some research questions as independent queries or to fix typos. In this case, you will be asked to approve our alterations. (In rare cases, queries had to be dropped, e.g., because they lead to no hits in our collection or were too similar to other queries.)

If you have any difficulties or questions, please do not hesitate to contact me.

Thanks again for your time and I look forward to your response,

Anna Ritchie

****

INSTRUCTIONS FOR AUTHORS:
APPROVING QUERY REFORMULATIONS AND MAKING RELEVANCE ASSESSMENTS

The purpose of this experiment is to expand on the relevance information you have already provided for a test collection for the ACL Anthology. You will be asked to approve our reformulation of your research questions as independent search queries and, for each question, to assess how relevant each of a list of papers is in relation to that question.

1) Research Questions

Read over the research question(s) from your paper, that you returned in the first stage. If a reformulation is given, decide whether you think this is an adequate representation of your original question as a stand-alone search query. If so, enter "yes" and proceed to step 2. Otherwise, please enter an appropriate reformulation of the question and return it to us. We will then re-enter it into the pooling process and generate a new list of papers for your relevance assessment.

2) Relevance Assessments

For each paper listed in your response form, decide whether that paper is relevant or irrelevant to the corresponding research question. Does that paper have some reasonable degree of relevance to that question i.e. would that paper be useful to someone trying to answer that question? Each paper is described by author and title in a web page, with a link to the full paper in PDF. The URL for this web page is given in your response form.

Please note that the list is randomized; relevant papers may appear below irrelevant ones in the list.

****

AUTHOR RESPONSE FORM

Research Question: How can we handle the problem of automatic identification of sources of opinions?
Reformulation: automatic identification of sources of opinions?
(Optional) reformulation:

Papers for Relevance Assessment:

```
(Paper details and PDFs available at
http://www.cl.cam.ac.uk/~ar283/H05-1045.choi.0.html)

PaperID         Relevant? ('X'=relevant, '0'=irrelevant)

I05-2030
P05-2006
J00-3003
H05-1044
W03-1017
H05-1043
H05-1116
H93-1031
N03-4017
P80-1020
W03-0404
H05-2017
W03-0613
I05-2011
P05-1015

**************************************************

Research Question: how can we model the problem with conditional random
fields?
Reformulation:  how can we model automatic identification of sources of
opinions with conditional random fields?
(Optional) reformulation:

Papers for Relevance Assessment:

(Paper details and PDFs available at
http://www.cl.cam.ac.uk/~ar283/H05-1045.choi.1.html)

PaperID         Relevant? ('X'=relevant, '0'=irrelevant)

320_pdf_2-col
W05-0622
W03-1017
P05-1056
21
176_Paper
I05-3027
I05-2030
W03-0430
H05-2019
X96-1027
P05-1044
N03-1028
I05-2011
200-771
```

B.2.4 Author response form (Phase Two)

```
AUTHOR RESPONSE FORM

Research Question: How can we handle the problem of automatic identification
of sources of opinions?
Reformulation:  automatic identification of sources of opinions?
==> YES
(Optional) reformulation:

Papers for Relevance Assessment:

(Paper details and PDFs available at
http://www.cl.cam.ac.uk/~ar283/H05-1045.choi.0.html)

PaperID         Relevant? ('X'=relevant, '0'=irrelevant)

I05-2030        x
```

```
P05-2006        x
J00-3003        0
H05-1044        x
W03-1017        x
H05-1043        x
H05-1116        x
H93-1031        0
N03-4017        x
P80-1020        0
W03-0404        x
H05-2017        x
W03-0613        0
I05-2011        x
P05-1015        0

**************************************************

Research Question: how can we model the problem with conditional random
fields?
Reformulation:  how can we model automatic identification of sources of
opinions with conditional random fields?
==> YES
(Optional) reformulation:

Papers for Relevance Assessment:

(Paper details and PDFs available at
http://www.cl.cam.ac.uk/~ar283/H05-1045.choi.1.html)

PaperID         Relevant? (‘X’=relevant, ‘0’=irrelevant)

320_pdf_2-col   0
W05-0622        x
W03-1017        0
P05-1056        0
21              x
176_Paper       x
I05-3027        0
I05-2030        0
W03-0430        x
H05-2019        x
X96-1027        0
P05-1044        0
N03-1028        x
I05-2011        0
200-771         0 =====> this is very much relevant to the first research
question instead.
```

B.3 Query reformulations

(Overleaf.)

Query ID	Reason	Original	Reformulation
H05-1003.ji.0	Filler	We present a novel mechanism for improving reference resolution by using the output of a relation tagger to rescore coreference hypotheses.	improving reference resolution by using the output of a relation tagger to rescore coreference hypotheses.
H05-1005.siddarthan.2	Anaphor, Typo	To use language modellling to predict the most plausible realization of the information, using the aligned strings	To use language modeling to predict the most plausible realization of information that is common across documents, using the aligned strings
H05-1011.moore.1	Anaphor, Context	What features should be included in the model?	What features should be included in a weighted linear model? - word alignment
H05-1011.moore.2	Anaphor	How can the best alignment according to the model be found?	How can the best word-alignment according to the weighted linear model be found?
H05-1011.moore.3	Anaphor, Context	How can the model weights be optimized?	How can the linear model weights be optimized? - word alignment
H05-1011.moore.4	Anaphor	How can the model be evaluated?	How can the weighted linear word alignment model be evaluated?
H05-1032.nomoto.0	Filler	The main issue the paper addresses is whether positional preferences the user may have in creating a summary can be exploited to possibly improve summarization.	whether positional preferences the user may have in creating a summary can be exploited to possibly improve summarization.
H05-1033.sporleder.2	Anaphor	Which machine-learning set-up is best suited for the task (i.e.classifier stacking, one-step vs. two-step chunking).	Which machine-learning set-up is best suited for discourse sentence chunking (i.e. classifier stacking, one-step vs. two-step chunking).
H05-1039.peng.0	Filler	The paper tries to combine linguistic analysis with surface pattern learning for definitional QA. It tries to understand the value of patterns and demonstrates the effectiveness of patterns for biographical questions.	combine linguistic analysis with surface pattern learning for definitional QA. understand the value of patterns and demonstrates the effectiveness of patterns for biographical questions.
H05-1039.peng.2	Context	how much labeling is required to obtain enough patterns?	how much labeling is required to obtain enough patterns? - question answering

H05-1042.lapata.2	Context	Identifying an appropriate domain	Identifying an appropriate domain - natural language generation
H05-1045.choi.0	Filler	How can we handle the problem of automatic identification of sources of opinions?	automatic identification of sources of opinions?
H05-1045.choi.1	Anaphor	how can we model the problem with conditional random fields?	how can we model automatic identification of sources of opinions with conditional random fields?
H05-1046.garbin.0	Filler	Given that a toponym(place name) can potentially refer to multiple places in news, can we use gazetteers and corpora to disambiguate different types of places, given scarce annotated data?	Given that a toponym (place name) can potentially refer to multiple places in news, can we use gazetteers and corpora to disambiguate different types of places, given scarce annotated data?
H05-1046.garbin.1	Filler	quantification of toponym ambiguity (more general than main)	quantification of toponym ambiguity
H05-1046.garbin.2	Filler	acquiring gazetteer resources (sub-problem but also more general)	acquiring gazetteer resources
H05-1046.garbin.3	Filler	finding alternatives to training a classifier on human-annotated data (sub-problem but also more general)	finding alternatives to training a classifier on human-annotated data
H05-1046.garbin.4	Filler	identifying features for toponym classifier (sub-problem but also more general)	identifying features for toponym classifier
H05-1046.garbin.5	Filler	evaluating the toponym classifier (sub-problem but also more general)	evaluating the toponym classifier
H05-1053.mccarthy.2	Context	automatic identification of words which would benefit from an automatically acquired first sense heuristic	automatic identification of words which would benefit from an automatically acquired first sense heuristic - word sense disambiguation
H05-1054.wu.0	Filler	The task of my paper is to find the optimal machine learning approach to identify the chinese Named Entities like person name, location name, organization name in Chinese Text.	find the optimal machine learning approach to identify the Chinese Named Entities like person name, location name, organization name in Chinese Text.

H05-1054.wu.2	Context	Search space is very large when only using statistical model, so we try to restrict the candidate generation by using human knowledge.	Search space is very large when only using statistical model, so we try to restrict the candidate generation by using human knowledge. - Chinese named entity recognition
H05-1057.raghavan.0	Filler	This paper aims to match inconsistently spelled names in ASR text, for example Lewinskey and Lewinski in order to boost performance of information retrieval on spoken document collections.	match inconsistently spelled names in ASR text, for example Lewinskey and Lewinski in order to boost performance of information retrieval on spoken document collections.
H05-1062.favre.2	Typo	what is the impact in term of ASR and NER performance of a temporal mismatch between the corpora used to train and test the models and how can it be recovered by means of meta-data information ?	what is the impact in terms of ASR and NER performance of a temporal mismatch between the corpora used to train and test the models and how can it be recovered by means of meta-data information ?
H05-1069.wang.1	Context	What second language would be best suitable to carry out the acquisition of sense examples?	What second language would be best suitable to carry out the acquisition of sense examples? - word sense disambiguation
H05-1069.wang.2	Typo	Given a second language, what resources would be best to use for acquiring sense examples, ie. what bilingual dictioanries and what monolingual corpora can achieve best WSD performance?	Given a second language, what resources would be best to use for acquiring sense examples, ie. what bilingual dictionaries and what monolingual corpora can achieve best WSD performance?
H05-1069.wang.3	Context	Given a set of sense examples, what machine learning algorithm can achieve high performance on this particular training data and why?	Given a set of sense examples, what machine learning algorithm can achieve high performance on this particular training data and why? - word sense disambiguation

H05-1073.alm.0	Filler	Short version: Is it possible to predict emotion and non-emotion from text? More explicit question statement: Since expressive text-to-speech synthesis would benefit from text-based emotion prediction, with what degree of accuracy is it possible to predict emotion and non-emotion, as well as emotional polarity, from text with a machine learning approach, given a modest corpus size and sentence-level classification?	Is it possible to predict emotion and non-emotion from text?
H05-1073.alm.1	Context	Does adding more sophisticated features, compared to content BOW or prior probability baseline, improve performance?	Does adding more sophisticated features, compared to content BOW or prior probab ility baseline, improve performance? - predict emotion
H05-1075.feng.0	Typo	Handling biograpical questions with implicature in a question answering system.	Handling biographical questions with implicature in a question answering system.
H05-1115.otterbacher.1	Anaphor	How can we use a graph-based approach in this problem?	How can we use a graph-based approach in question-focused sentence retrieval?
H05-1115.otterbacher.3	Anaphor	How can we perform the above tasks given the prevalence of paraphrasing in news texts?	How can we perform question-focused sentence retrieval and automatic answer finding given the prevalence of paraphrasing in news texts?
P05-1005.kohomban.1	Anaphor	if we can do the above, how can we learn those concepts from a generic set of labelled data, overcoming noise?	if we can generalize learning word senses, how can we learn those concepts from a generic set of labelled data, overcoming noise?

B.4 Queries

Query ID	Query
H05-1001.steinberger.0	Does anaphora resolution improve summarization (based on latent semantic analysis) performance?
H05-1001.steinberger.1	Using anaphora (coreference) resolution and its applications.
H05-1001.steinberger.3	Evaluation of text summarization and anaphora resolution.
H05-1005.siddarthan.3	problems with extractive multilingual summarization
H05-1015.filatova.0	How to learn occupation-related activities? Is the classification of people according to their occupations based on automatically learned occupation-related activities reliable?
H05-1015.filatova.1	Divide the biography into 3 parts containing the following types of activities: general biographical; occupation-related; person-specific
H05-1015.filatova.2	Automatically extract candidate activities used for the description of people of various occupations
H05-1022.deng.0	how to build word alignment models with high quality alignments and efficient training algorithms for statistical machine translation?
H05-1022.deng.1	How to induce statistical phrase translation models from word alignments and their model?
H05-1032.nomoto.0	whether positional preferences the user may have in creating a summary can be exploited to possibly improve summarization.
H05-1033.sporleder.0	Is it possible to borrow ideas from syntactic chunking to automatically determine the elementary discourse units and their functions in sentences and is this useful for sentence compression.
H05-1033.sporleder.1	Can knowledge-lean methods be used to discourse chunk a sentence?
H05-1033.sporleder.2	Which machine-learning set-up is best suited for discourse sentence chunking (i.e. classifier stacking, one-step vs. two-step chunking).
H05-1033.sporleder.3	Can sentences be compressed by discourse chunking them automatically and then dropping all satellite spans?
H05-1045.choi.0	automatic identification of sources of opinions?
H05-1045.choi.1	how can we model automatic identification of sources of opinions with conditional random fields?

H05-1046.garbin.0	Given that a toponym (place name) can potentially refer to multiple places in news, can we use gazetteers and corpora to disambiguate different types of places, given scarce annotated data?
H05-1046.garbin.1	quantification of toponym ambiguity
H05-1046.garbin.2	acquiring gazetteer resources
H05-1046.garbin.3	finding alternatives to training a classifier on human-annotated data
H05-1046.garbin.4	identifying textual features for a toponym classifier
H05-1046.garbin.5	evaluating the toponym classifier
H05-1053.mccarthy.0	Establishing that an automatic method to acquire predominant senses can outperform a manually derived first sense heuristic when dealing with WSD of domain specific text for certain types of words.
H05-1053.mccarthy.1	The production of sense-annotated domain specific corpora for evaluation
H05-1053.mccarthy.2	automatic identification of words which would benefit from an automatically acquired first sense heuristic - word sense disambiguation
H05-1054.wu.0	Find the optimal machine learning approach to identify the Chinese Named Entities like person name, location name, organization name in Chinese Text.
H05-1054.wu.1	For the current Word Model for Chinese Named Entity Recognition, data sparseness problem is very serious. Therefore, we want to find a solution to resolve it.
H05-1054.wu.2	Search space is very large when only using statistical model, so we try to restrict the candidate generation by using human knowledge. - Chinese named entity recognition
H05-1057.raghavan.0	match inconsistently spelled names in ASR text, for example Lewinskey and Lewinski in order to boost performance of information retrieval on spoken document collections.
H05-1059.tsuruoka.0	How can we fully utilize information about tag sequences in machine-learning based algorithms for sequence tagging tasks?
H05-1068.munson.0	How well does greedy ensemble selection optimize difficult and cumbersome performance metrics for natural language processing problems?
H05-1068.munson.2	Can a computer find noun phrase coreference chains in a document?
H05-1068.munson.3	Can a computer automatically identify words in a document that express perspective, opinion, or private state?
H05-1068.munson.4	Given perspective, opinion, and private state words, can a computer infer the hierarchy among the different perspectives?

H05-1069.wang.0	Would monolingual corpora in a second language, such as Chinese, be a good resource to obtain training data (sense examples) for machine learning Word Sense Disambiguation (WSD) systems?
H05-1069.wang.1	What second language would be best suitable to carry out the acquisition of sense examples? - word sense disambiguation
H05-1069.wang.2	Given a second language, what resources would be best to use for acquiring sense examples, ie. what bilingual dictionaries and what monolingual corpora can achieve best WSD performance?
H05-1069.wang.3	Given a set of sense examples, what machine learning algorithm can achieve high performance on this particular training data and why? - word sense disambiguation
H05-1073.alm.0	Is it possible to predict emotion and non-emotion from text?
H05-1073.alm.1	Can more sophisticated features benefit emotion prediction?
H05-1075.feng.0	Handling biographical questions with implicature in a question answering system.
H05-1078.merlo.0	Can we build a parser that outputs semantic role annotation?
H05-1078.merlo.1	Does learning semantic roles improve parsing performance?
H05-1079.markert.0	How well can classical inference engines, namely theorem proving and model building, be adapted for solving the textual entailment problem?
H05-1081.surdeanu.0	Do combination strategies improve semantic role labeling?
H05-1081.surdeanu.1	What is the state-of-the-art on semantic role labeling using real syntax?
H05-1091.bunescu.0	Given a document containing noun phrases annotated with predefined types of entities (such as Person, Organization, Location, Facility, and Geo-Political Entity), where are the instances where the text asserts a relationship (such as Role, Located At, Near, Social) between pairs of entities?
H05-1091.bunescu.1	What is the word-word dependency structure of a sentence?

H05-1096.ueffing.0	How can we automatically calculate measures of confidence for single words in machine translation output?
H05-1096.ueffing.1	How can the information given in state-of-the-art models for statistical machine translation be explored for confidence estimation?
H05-1107.hwa.0	How can we leverage from multiple sources of information to acquire annotated resources for training a Chinese Part-of-Speech tagger.
H05-1107.hwa.1	Active learning – if we can only afford to annotate a small amount of Chinese data, what kind of data should be annotated so as to be the most helpful in training the tagging model?
H05-1107.hwa.2	Projecting resources – can we take advantage of high quality tagged data for English and the availability of parallel corpus to produce automatically tagged Chinese data (to train a Chinese tagger)?
H05-1107.hwa.3	Combining information sources – what is the best way to combine the model trained from the small manually annotated data and the model trained from projected data.
H05-1108.pado.0	How can role-semantic information be transferred between parallel sentences in different languages?
H05-1108.pado.1	How to assign a role-semantic analysis to a sentence ("Shallow semantic parsing")
H05-1108.pado.2	The usefulness of role-semantic analyses for NLP tasks
H05-1111.swier.0	What is the benefit, if any, of exploiting knowledge contained in verb lexicons for the task of automatically labelling semantic roles?
H05-1111.swier.1	To what degree is it possible to adapt, via a role mapping, a corpus annotated with a fine-grained set of semantic roles for the purpose of evaluating a role labelling system that uses a coarser grained role set?
H05-1115.otterbacher.1	How can we use a graph-based approach in question-focused sentence retrieval?
H05-1115.otterbacher.3	How can we perform question-focused sentence retrieval and automatic answer finding given the prevalence of paraphrasing in news texts?
H05-1122.olney.0	Is it possible to use existing methods for monologue topic segmentation on tutorial dialogue
P05-1002.osborne.0	How can CRFs be made to scale with large numbers of labels?
P05-1002.osborne.1	Is it possible to select a highly informative number of bits when creating error-correcting codes?

P05-1003.osborne.0	How can CRFs be regularised without using parameterised priors?
P05-1003.osborne.1	How are LOP-CRFs trained?
P05-1003.osborne.2	How do LOP-CRFs compare with regularised CRFs?
P05-1003.osborne.3	Is diversity important for good logarithmic opinion pool conditional random field performance?
P05-1005.kohomban.0	Can we generalize learning word senses by using a common set of super-senses instead of an enumerative lexicon?
P05-1009.soricut.0	How to perform the intersection of IDL-expressions with n-gram language models?
P05-1009.soricut.1	How to perform natural language generation for text-to-text applications?
P05-1013.nivre.0	How can non-projective dependencies be captured accurately and efficiently in dependency-based syntactic parsing?
P05-1013.nivre.1	Can non-projective dependencies be captured with an accuracy sufficient to improve overall parsing accuracy?
P05-1013.nivre.2	Can non-projective dependencies be captured with an accuracy sufficient to outperform the best projective dependency parsers?
P05-1021.yang.0	How to effectively utilize statistics-based semantic compatibility information to improve pronoun resolution.
P05-1030.rieser.1	What clarification classification scheme is suited to describe naturally occurring CRs in order to generate them?
P05-1031.schlangen.0	Can Fragments, a certain class of non-sentential utterances, be automatically detected and linked up with their antecedents, and can criteria for this task be learned using machine learning techniques?
P05-1031.schlangen.1	Can the class of non-sentential utterances that do have an individual antecedent be consistently defined?
P05-1035.amigo.0	Automatic evaluation of summaries and automatic metaevaluation of metrics. How to combine and meta-evaluate similarity metrics to measure the proximity from an automatic summary to a set of models.
P05-1035.amigo.1	Combining metrics and similarities from models (manual summaries) without considering metric scales.
P05-1035.amigo.2	Defining criteria for the meta-evaluation of metrics; human judges are expensive.
P05-1035.amigo.3	Defining a measure to estimate the reliability of the set of evaluated summaries which have been used to meta-evaluate metrics. That is, are the automatic summaries in the corpus representative from the possible automatic solutions?

B.5 Phase Two pooling Lemur parameters

Model	Parameter	Setting
Okapi	BM25K1	1.2
	BM25B	0.75
	BM25K3	7
	BM25QTF	0.5
	feedbackTermCount	20
KL	smoothStrategy	jm
	JelinekMercerLambda	0.5
	DirichletPrior	2000
	feedbackTermCount	20
	queryUpdateMethod	0
	feedbackCoefficient	0.5
	feedbackProbThresh	0.001
	feedbackProbSumThresh	1
	feedbackMixtureNoise	0.5
	emIterations	50

Appendix C

Citation and reference grammar

Regular expressions for citations

```
our $SURNAME = '()(<SURNAME>[^<]*?</SURNAME>)()';
# We can't use this 'perfect' expression (below) for efficiency reasons
# It makes sure there's never another <SURNAME> between the first and the </SURNAME>
#our $SURNAME = '()(<SURNAME>([^<]*(<(?!SURNAME))?[^<]*)+</SURNAME>)()';

our $GENITIVE = "()('(\\s)*s)()";
our $PRENAME = '(?:\b)(?:[Ll]a|[Dd][iue]|[dD]ella|[Dd]e\s+[Ll]a|[Vv]an\s*?(?:[td]e[rn])?)(?:)';
our $NAME = '()((' . $PRENAME . '(\s*))?' . $SURNAME . ')()';
our $NAMECOMMA = '()(' . $NAME . '(\s)*?(' . $COMMA . ')?)((\s)*)';
our $NAMESANDNAME = '()((' . $NAMECOMMA . ')+(\s)*' . $AND . '(\s)*' . $NAME . ')()';
our $NAMES = '()(' . $NAMESANDNAME . '|' . $NAME . ')((\s)*?(' . $GENITIVE . ')?)';

our $ETAL = '(\s+)([Ee][tT](?:\.)?(?:\s)*?[Aa][lL](?:\.)?)(\s*)';
our $NAMEETAL =
  '()(' . $NAME . '(\s)*?(' . $COMMA . ')?(\s)*?' . $ETAL . '(\s)*?(' . $FULLSTOP . ')?)((\s)*?(' . $GENITIVE . ')?)';
our $AUTHOR = '()(' . $NAMEETAL . '|' . $NAMES . ')()';

# dates can include distinguishing letters
our $DATELETTERS = '(?:)(?:[abcdefghijklmnopqs])(?:)';
# years can be 18**, 19**, or 20** or just ** (two digits)
our $YEAR = '()(((?<!\d)(19|20|18)\d{2})|(\d{2})(?!\d))()';
# restrict year dates to having a single dateleter  (no space, no comma)
our $SIMPLEDATENUMS = '()(' . $YEAR . '(' . $DATELETTERS . ')?)()';
our $DATEWORDS =
  '(?:)(?:(?:to appear)|(?:forthcoming)|(?:in press)|(?:in print)|(?:in preparation)|
  (?<=\b)(?:forth(?:(?:\.|coming)?))|(?:submitted)|(?:this issue))(?:)';
our $SIMPLEDATE = '()((?<=\b)' . $SIMPLEDATENUMS . '|' . $DATEWORDS . ')()';
our $SIMPLEDATECOMMA = '()(' . $SIMPLEDATE . '((\s)*' . $COMMA . ')?)(\s*)';
our $SIMPLEDATEORLIST = '()((' . $SIMPLEDATECOMMA . ')+((\s)*' . $AND . ' (\s)*' . $SIMPLEDATE. ')?)()';

our $RANGE = '()(\d+\s*' . $DASH . '\s*\d+)()';
our $EXTENTNUM = '()(' . $RANGE . '|(\d+)\.(\d+)|(\d+)ff(\.)?|(\d+)(?=\b)(?!\S*\<\/DATE>))()';
our $EXTENTWORDS =
  '()(([Cc]hapter|[Pp]age|[Pp]g(\.)?|[Pp]p?(\.)?|[cC]h(a?)(p?)(t?)(r?))[s\.]?|[Ss]ection|[Ss](\.?))()';
# NB we need the double brackets round the ?<= expression to keep the $1 etc variables happy
our $EXTENT = '((?<=\b))((' . $EXTENTWORDS . '(\s*))?' . $EXTENTNUM . ')()';
our $EXTENTCOMMA = '()(' . $EXTENT . '((\s)*' . $COMMA . ')?)((\s)*)';
our $EXTENTORLIST = '()((' . $EXTENTCOMMA . ')+((\s)*' . $AND . ' (\s)*' . $EXTENT. ')?)()';

# Need to deal with definitions of acronyms/abbreviations
our $ACRONYMCHAR = '([A-Z1-9]|(' . $AND . '\s?))';
our $ACRONYM = '()([A-Z]' . $ACRONYMCHAR . '+)()';
our $ACRONYMQUOTES = '()(' . $QUOTE . '\s*)?' . $ACRONYM . '(\s*' . $QUOTE . ')?)()';
our $HENCEFORTHWORD = '()([Hh]ence-?\s?forth)()';
our $HENCEFORTHFIRST = '()(' . $HENCEFORTHWORD . '\s*(' . $COMMA . ')?\s*(' . $ACRONYMQUOTES . ')()';
our $HENCEFORTHLAST = '()(' . $ACRONYMQUOTES . '\s*(' . $COMMA . ')?\s*(' . $HENCEFORTHWORD . ')()';
our $HENCEFORTHFIRSTORLAST = '()((' . $HENCEFORTHFIRST . '|' . $HENCEFORTHLAST . '))()';
our $BRACKETEDHENCEFORTH = '(' . '$LBR\s*)(' . $HENCEFORTHFIRSTORLAST. ')(\s*' . $RBR  . ')';
```

```
our $UNBRACKETEDHENCEFORTH = '()(' . $HENCEFORTHFIRSTORLAST. ')()';
our $HENCEFORTH = '()(' . $BRACKETEDHENCEFORTH . '|' . $UNBRACKETEDHENCEFORTH . ')()';

# It's possible to have more than one extent per title
our $SIMPLEDATEEXTENT = '()(' . $SIMPLEDATEORLIST . '((\s*' . $COMMA . ')?\s*' . $EXTENTORLIST . ')?)()';
our $SIMPLEDATEEXTENTCOMMA = '()(' . $SIMPLEDATEEXTENT . '(\s)*?(' . $COMMA . ')?)((\s)*)';
our $SIMPLEDATEEXTENTLIST = '()(((. $SIMPLEDATEEXTENTCOMMA . ')+(\s)*' . $AND . ' (\s)*' . $SIMPLEDATEEXTENT. ')()';
our $SIMPLEDATEEXTENTORLIST = '()(' . $SIMPLEDATEEXTENTLIST. '|' . $SIMPLEDATEEXTENT . ')()';
our $UNBRACKETEDDATE = '()(' . $SIMPLEDATEEXTENTORLIST . ')()';
our $BRACKETEDDATE = '()(' . $LBR . '(\s)*(' . $UNBRACKETEDDATE . ')+(\s)*?' . $RBR . ')()';
our $DATE = '()(' . $BRACKETEDDATE . '|' . $UNBRACKETEDDATE . ')()';
our $BRACKETEDDATEHENCEFORTH =
  '(' . $LBR  . '\s*)(' . $DATE . '((\s*' . $COMMA . ')?\s*' . $HENCEFORTH . ')?)(\s*' . $RBR  . ')';
our $UNBRACKETEDDATEHENCEFORTH  = '()(' . $DATE . '((\s*' . $COMMA . ')?\s*' . $HENCEFORTH . ')?)()';
our $DATEHENCEFORTH = '()(' . $BRACKETEDDATEHENCEFORTH . '|' . $UNBRACKETEDDATEHENCEFORTH . ')()';
our $PC = '()((p\.?c\.?)|(personal communication))()';
our $DATEHENCEFORTHPC = '()(' . $DATEHENCEFORTH . '|' . $PC . ')()';

our $AUTHORSIMPLEDATE = '()(' . $AUTHOR . '(\s)*(' . $COMMA . ')?(\s)*(' . $DATEHENCEFORTHPC . ')+)()';
our $AUTHORSIMPLEDATECOMMA =
  '()(' . $AUTHORSIMPLEDATE . '(\s)*(' . $COMMA . '|((' . $COMMA . '(\s)*)?'  . $AND . '))?(\s)*)()';

our $PRESTRINGWORD =
  '(\b)((see\b)|(also\b)|(e(\.)?(\s)*g(\.)?)|(in\b)|(for example\b)|(such as\b)|([cC](\.)?f(\.)?))()';
our $PRESTRING = '()(' . $PRESTRINGWORD . '(((\s*' . $COMMA. ')?\s*' .$PRESTRINGWORD . ')+)?(\s*' . $COMMA. ')?)()';
our $POSTSTRING = '()((inter\s?alia)|(among\s+others))()';




our $PARENTHETIC =
  '()(' . $LBR . '(\s)*(' . $PRESTRING . ')?(\s)*('  . $AUTHORSIMPLEDATECOMMA . ')+(\s)*(' . $POSTSTRING .
  ')?(\s)*' . $RBR . ')()';
our $SYNTACTIC =
  '()(' . $AUTHOR . '(\s)*(' . $COMMA . ')?(\s)*(' . $DATEHENCEFORTHPC . ')((\s)*' . $POSTSTRING . ')?)()';

# For finding sequences of REFAUTHORs
our $OPENREFAUTHOR = '<REFAUTHOR[^>]+>';
our $CLOSEREFAUTHOR = '</REFAUTHOR>';
```

Regular expressions for references

```
our $REFLIST_YEAR = '(?:(?<!\d)(?:19|20|18)\d{2})';
our $REFLIST_DATENUMS = '(?:' . $REFLIST_YEAR . '(?:' . $DATELETTERS . ')?)';
our $REFLIST_DATEWORDS = '(?:' . $REFLIST_DATENUMS . '\s*)?' . $DATEWORDS . ')';
our $REFLIST_DATE = '(' . $LBR . '?)(' . $REFLIST_DATENUMS . '|' . $DATEWORDS . ')(' . $RBR . '?)';

our $INITIALBASIC = '[A-Z](?:\.)?';
our $INITIALHYPHEN = $INITIALBASIC . '-' . $INITIALBASIC;
our $INITIAL = '(?:' . $INITIALBASIC . '|' . $INITIALHYPHEN .  ')';
# We don't necessarily need whitespace between initials in a list
our $INITIALSEQ = '(?:(?:' . $INITIAL . '\s*)?' . $INITIAL . ')';

# Author names (lists of)
our $SURNAME_MAC_OR_O = '(?:(?:Ma?c)|(?:O\'))';
our $REFLIST_SURNAMESIMPLE =  '(?:[A-Z][a-z\']+)';
our $REFLIST_SURNAMEMAC = '(?:' . $SURNAME_MAC_OR_O . '?' . $REFLIST_SURNAMESIMPLE . ')';
our $REFLIST_SURNAMEHYPHEN = '(?:' . $REFLIST_SURNAMEMAC . '-'. $REFLIST_SURNAMEMAC. ')';
our $REFLIST_SURNAMEBASIC = '(?:' . $REFLIST_SURNAMEHYPHEN .'|' . $REFLIST_SURNAMEMAC . ')';
our $REFLIST_SURNAME = '(?:'. $PRENAME. '\s+)?' . $REFLIST_SURNAMEBASIC . '(?:,\s(?:(?:(?:Jr|Sr)\.)|(?:II(?:I)?)))?';
our $REFLISTNAMEBASIC = '(?:[A-Z][a-z\']+)';
our $REFLISTNAMEHYPHENSECOND = '(?:[A-Z]?[a-z\']+)';
our $REFLISTNAMEHYPHEN = '(?:' . $REFLISTNAMEBASIC . '-' . $REFLISTNAMEHYPHENSECOND .')' ;
our $REFLISTNAME = '(?:' . $REFLISTNAMEHYPHEN . '|' . $REFLISTNAMEBASIC .')' ;
our $REFLIST_FIRSTNAMESPART = '(' . $REFLISTNAME . '|' . $INITIALSEQ. ')';
our $REFLIST_FIRSTNAMES = '(' . $INITIALSEQ .  ')';
our $REFLIST_FULLNAME = '(' . $REFLIST_SURNAME . ',\s+' . $REFLIST_FIRSTNAMES .')';
```

```
# We need whitespace between names in a list
our $NAMESEQ = '(?:(?:' . $REFLISTNAME . '|' . $INITIALSEQ . ')\s+)*(?:' . $REFLISTNAME . '|' . $INITIALSEQ . ')';

our $ONE_AUTHORLIST_SEP = '(?:' . $COMMA . '|' . $AND . ')';
our $OXFORD_AUTHORLIST_SEP = '(?:' . $COMMA . '\s*' . $AND . ')';
our $AUTHORLIST_SEP = '(?:' . $OXFORD_AUTHORLIST_SEP . '|' . $ONE_AUTHORLIST_SEP . ')';

our $SURNAME_COMMA_INITIALS = '(' . $REFLIST_SURNAME . ')(\s*,\s*' . $INITIALSEQ . ')';
our $SURNAME_COMMA_NAMES = '(' . $REFLIST_SURNAME . ')(\s*,\s*' . $NAMESEQ . ')';
our $INITIALS_SURNAME = '(' . $INITIALSEQ . '\s*' . $REFLIST_SURNAME . ')';
our $NAMES_SURNAME = '(' . $NAMESEQ . '\s+)(' . $REFLIST_SURNAME . ')';
# We need whitespace to separate surname and initials when there's no comma
our $SURNAME_INITIALS = '(' . $REFLIST_SURNAME . ')(\s+' . $INITIALSEQ . ')';
```

Regular expressions for Anthology publication names

```
our $JOURNAL_PATTERN = '((Journal of )?Computational Linguistics)';
our $ACL_PATTERN = '((Proceedings of ACL)|((Annual Meeting)?.*Association (for|of) Computational Linguistics)|([^A-Za-z0-9]ACL[^A
our $COLING_PATTERN = '((COLING)|(International Conference on Computational Linguistics))';
our $HLT_PATTERN = '((Proceedings of HLT)|(Human Language Technology))';
our $NAACL_PATTERN = '((Proceedings of NAACL)|(North American Association (for|of) Computational Linguistics))';
our $EACL_PATTERN = '((EACL)|(European Chapter of the Association (for|of) Computational Linguistics))';
our $ANLP_PATTERN = '((ANLP)|(Applied Natural Language Processing))';
our $TIPSTER_PATTERN = '(TIPSTER)';
our $TINLAP_PATTERN = '(TINLAP)';
our $MUC_PATTERN = '(([^A-Za-z0-9]MUC[^A-Za-z])|(Message Understanding Conference))';
our $IJCNLP_PATTERN = '(IJCNLP)';
our $WORKSHOP_PATTERN =
  '((C[oO]NLL)|(Computational Natural Language Learning)|(EMNLP)|(Empirical Methods.*Natural Language Processing)|
  (SIGDAT)|(SIGDIAL)|(SIGLEX)|(SIGNLL)|(SIGGEN))|(SIGPHON)|(SIGHAN)|(SENSEVAL)|
  (Evaluation of Systems for the Semantic Analysis of Text)|(INLG)|
  (International Workshop on Natural Language Generation)|(VLC)|(Very Large Corpora)|
  (Reversible Grammar in Natural Language Processing)|(Lexical Semantics and Knowledge Representation)|
  (Acquisition of Lexical Knowledge from Text)|(Intentionality and Structure in Discourse Relations)|
  (Combining Symbolic and Statistical Approaches to Language)|(Computational Phonology)|
  (Breadth and Depth of Semantic Lexicons)|(Tagging Text with Lexical Semantics)|(Spoken Language Translation)|
  (Natural Language Processing for Communication Aids)|(Interactive Spoken Dialog Systems)|
  (Intelligent Scalable Text Summarization)|
  (Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications)|
  (From Research to Commercial Applications: Making NLP Work in Practice)|
  (Concept to Speech Generation Systems)|
  (Operational Factors in Practical,? Robust Anaphora Resolution for Unrestricted Texts)|
  (Referring Phenomena in a Multimedia Context and their Computational Treatment)|
  (Computational Environments for Grammar Development and Linguistic Engineering)|
  (Content Visualization and Intermedia Representations)|(CVIR)|(Discourse Relations and Discourse Markers)|
  (Processing of Dependency-Based Grammars)|(Computational Treatment of Nominals)|
  (Usage of WordNet in Natural Language Processing Systems)|
  (Partially Automated Techniques for Transcribing Naturally Occurring Continuous Speech)|
  (Computational Approaches to Semitic Languages)|(Finite State Methods in Natural Language Processing)|
  (Natural Language Generation)|(The Relation of Discourse/Dialogue Structure and Reference)|
  (Coreference and Its Applications)|(Towards Standards and Tools for Discourse Tagging)|
  (Computer Mediated Language Assessment and Evaluation in Natural Language Processing)|
  (Computer and Internet Supported Education in Language and Speech Technology)|
  (Unsupervised Learning in Natural Language Processing)|
  (Syntactic and Semantic Complexity in Natural Language Processing Systems)|(Applied Interlinguas)|
  (Conversational Systems)|(Automatic Summarization)|(Embedded Machine Translation Systems)|
  (Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Systems)|
  (Word Senses and Multi-linguality)|(Comparing Corpora)|(
  Recent Advances in Natural Language Processing and Information Retrieval)|
  (Chinese Language Processing Workshop)|(European Workshop on Natural Language Generation)|(EW?NLG)|
  (Evaluation Methodologies fro Language and Dialogue Systems)|(Human Language Technology and Knowledge Management)|
  (Open-Domain Question Answering)|(Temporal and Spatial Information Processing)|
  (Data-Driven Methods in Machine Translation)|(Sharing Tools and Resources)|
  (Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics)|
  (Natural Language Processing in( the)? Biomedic)|(Morphological and Phonological Learning)|
  (Speech-to-Speech Translation: Algorithms and Systems)|
  (Word Sense Disambiguation: Recent Successes and Future Directions)|(Unsupervised Lexical Acquisition)|
  (Building and Using Semantic Networks)|(SEMANET)|(Asian Language Resources and International Standardization)|
```

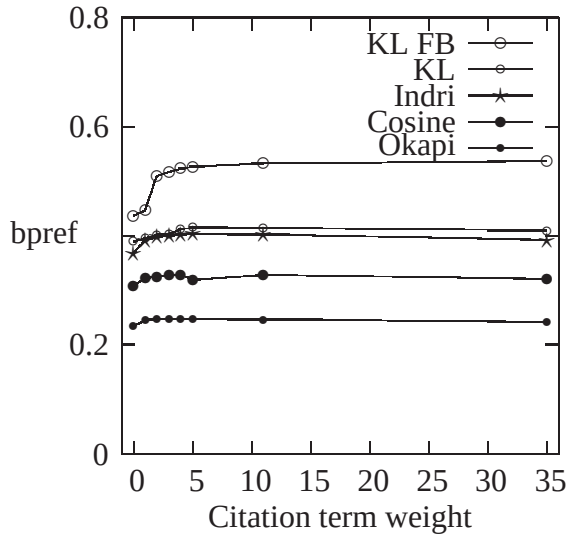
```
  (A Roadmap for Computational Linguistics)|(Computational Terminology)|(COMPUTERM)|(CompuTerm)|
  (Grammar Engineering and Evaluation)|(Machine Translation in Asia)|(NLP( and )?XML)|
  (Multilingual Summarization and Question Answering)|(Analysis of Geographic References)|
  (Building Educational Applications Using Natural Language Processing)|
  (Building and Using Parallel Texts: Data Driven Machine Translation and Beyond)|
  (Text Summarization( Branches Out)?)|(Learning Word Meaning from Non-Linguistic Data)|
  (Research Directions in Dialogue Processing)|(Software Engineering and Architecture of Language Technology Systems)|
  (SEALTS)|(Text Meaning)|(Information Retrieval with Asian Languages)|(Lexicon and Figurative Language)|
  (Multilingual and Mixed-language Named Entity Recognition)|(Paraphrasing)|(Multiword Expressions:)|
  (Linguistic Annotation: Getting the Model Right)|(Patent Corpus Processing)|
  (Current Themes in Computational Phonology and Morphology)|(Discourse Annotation)|
  (Incremental Parsing: Bringing Engineering and Cognition Together)|(Question Answering in Restricted Domains)|
  (Reference Resolution and Its Applications)|
  (Automatic Alignment and Extraction of Bilingual Domain Ontology for Medical Domain Web Search)|(NLPBA)|(BioNLP)|
  (Psycho-Computational Models of Human Language Acquisition)|
  (Language Resources for Translation Work,? Research and Training)|(Recent Advances in Dependency Grammar)|
  (Computational Approaches to Arabic Script-based Languages)|(eLearning for Computational Linguistics)|
  (Linguistically Interpreted Corpora)|(R[Oo]bust Methods in Analysis of Natural Language Data)|(ROMAND)|
  (Enhancing and Using Electronic Dictionaries)|(Multilingual (Linguistic|Language) Resources)|
  (Pragmatics of Question Answering)|(Computational Lexical Semantics)|
  (Frontiers in (Corpus Annotation|Linguistically Annotated Corpora))|(Scalable Natural Language Understanding)|
  (ScaNaLU)|(Interdisciplinary Approaches to Speech Indexing and Retrieval)|
  (Spoken Language Understanding for Conversational Systems and Higher Level Linguistic Information for Speech Processing)|
  (Linking Biological Literature,? Ontologies and Databases)|
  (Feature Engineering for Machine Learning in Natural Language Processing)|
  (Psychocomputational Models of Human Language Acquisition)|
  (Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization)|
  (Deep Lexical Acquisition)|(Empirical Modeling of Semantic Equivalence and Entailment)|
  (Parsing Technology)|(Information Extraction Beyond The Document)|(Sentiment and Subjectivity in Text)|
  (Constraints and Language Processing)(Ontology Learning and Population)|
  (Task-Focused Summarization and Question Answering)|
  (How Can Computational Linguistics Improve Information Retrieval)|(Annotating and Reasoning about Time and Events)|
  (Linguistic Distances)|(Tree Adjoining Grammar and Related Formalisms)|(Web as Corpus)|
  (Knowledge and Reasoning for Language Processing)|(KRAQ)|(Multilingual Question Answering)|(MLQA)|
  (Cross-Language Knowledge Induction)|(Prepositions)|(Adaptive Text Extraction and Mining)|(ATEM)|
  (Multi-word-expressions in a multilingual context)|(Making Sense of Sense)|
  (Learning Structured Information in Natural Language Applications)|
  (NEW TEXT Wikis and blogs and other dynamic text sources)|(Interactive Question Answering)|
  (Statistical Machine Translation)|(Linking Natural Language and Biology)|
  (Analyzing Conversations in Text and Speech)|
  (Computationally Hard Problems and Joint Inference in Speech and Language Processing)|
  (Medical Speech Translation)|(Graph Based Methods for Natural Language Processing)';
```

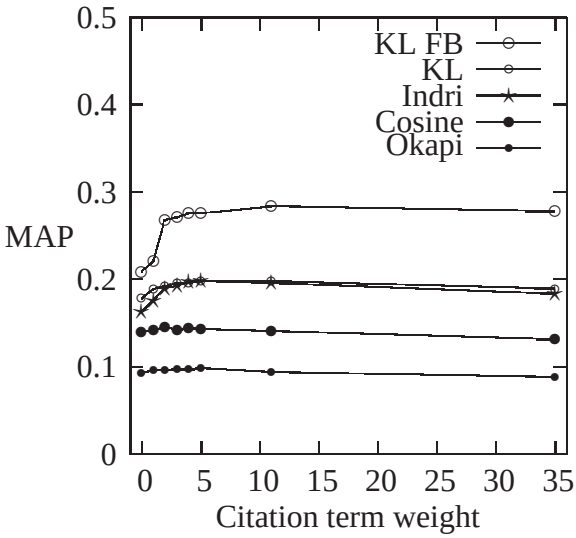
Miscellaneous regular expressions

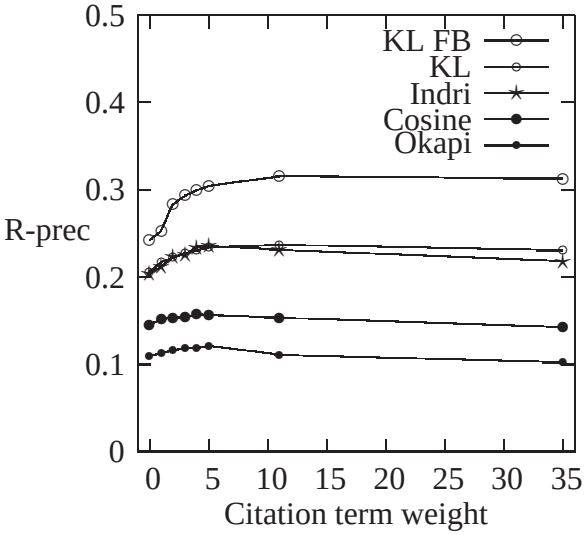
```
our $COMMA = "(?:[,;:])";
our $SLASH = '([\/\\\])';
our $PERCENT = "(%)";
our $FULLSTOP = '(\.)';
# We need the whitespace after the &amp;amp; since it appears that's what's added in place of &
our $AMPERSAND = '(?:(?:\&amp;amp;\s?)|\&)';
our $AND = '(?:' . $AMPERSAND . '|and)';
our $QUOTE = "(``|''|`|'|\")";
our $LBR = '(?:[\[{(])';
our $RBR = '(?:[\]}\)])';
our $DASH = '([-]+)';
our $QM = '(\?)';
our $EM = '(!)';
our $XMLQUOTE = '([\'"])';

our $MIXEDCASE = '([a-zA-Z]+)';
our $LOWERCASE = '([a-z]+)';
our $UPPERCASE = '([A-Z]+)';
our $DECIMAL = '(\d+)\.(\d+)';
our $SIMPLENUM = '(\d+)';

our $WORD = "$LOWERCASE|$UPPERCASE|$MIXEDCASE|$DECIMAL|$SIMPLENUM";
our $NOTWORD = "$COMMA|$SLASH|$PERCENT|$FULLSTOP|$AMPERSAND|$QUOTE|$LBR|$RBR|$DASH|$QM|$EM";
our $ALL = "$WORD|$NOTWORD";
```

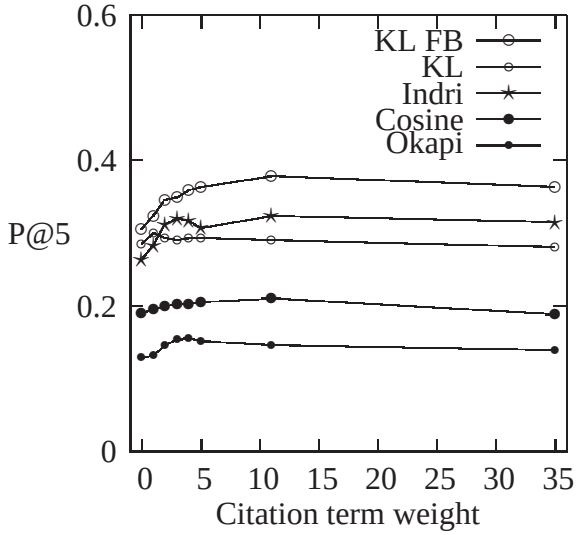
Appendix D

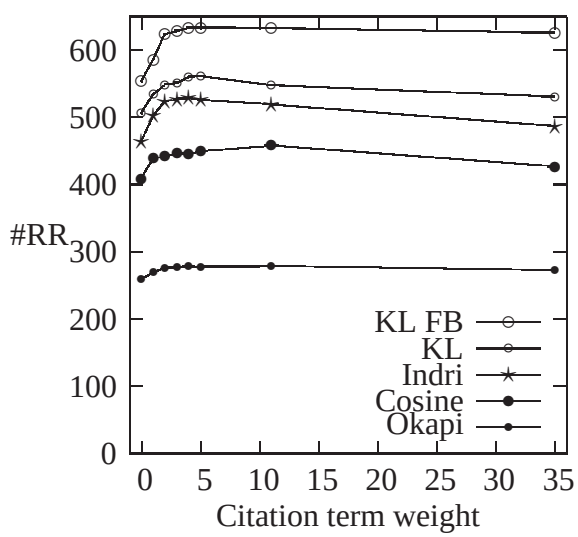
Plots of experimental results

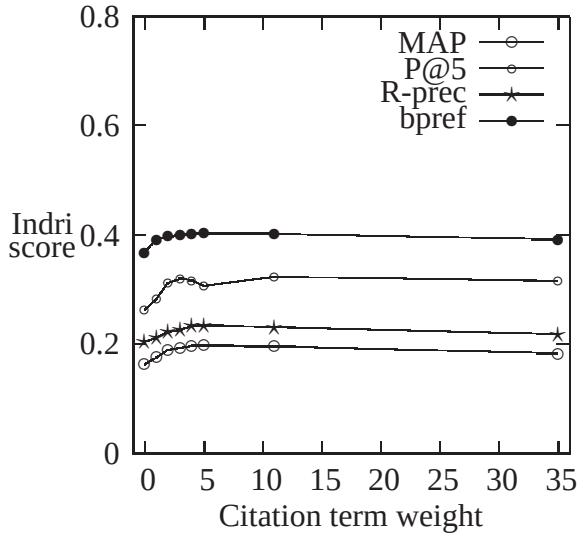


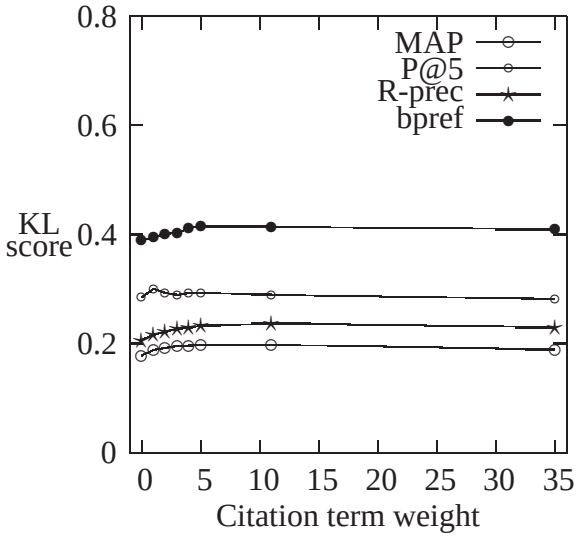


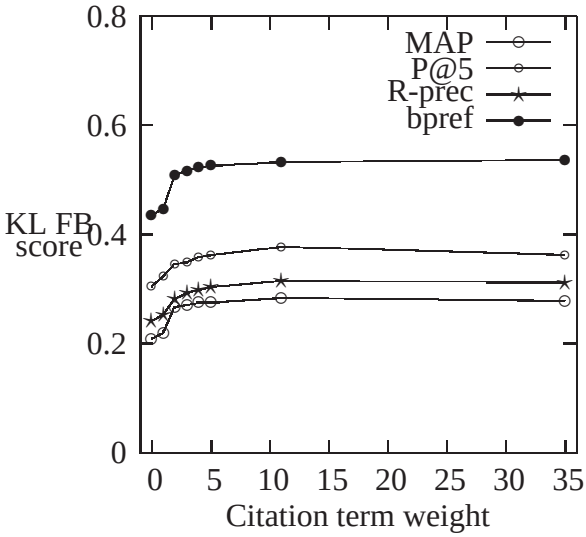


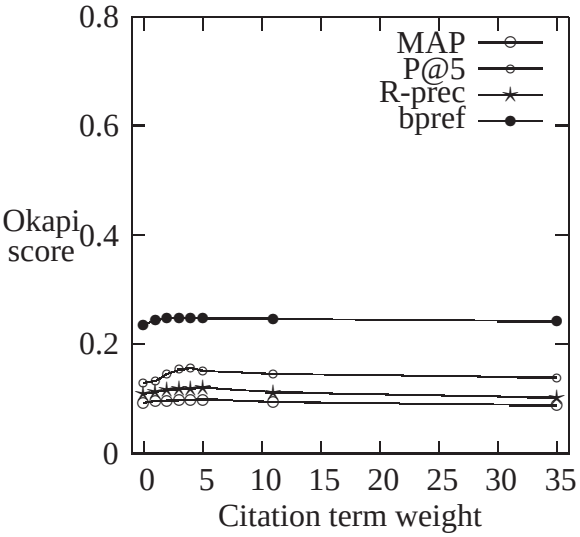


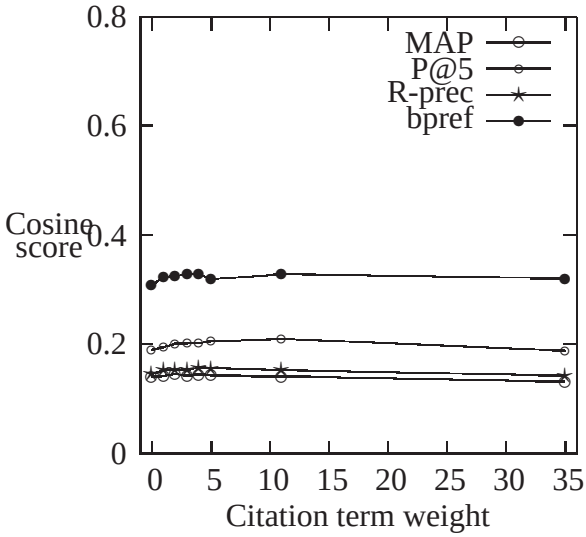

#RR
600
500
400
300
200
100
0
0 5 10 15 20 25 30 35
Citation term weight
KL FB
KL
Indri
Cosine
Okapi


Indri score
0.8
0.6
0.4
0.2
0
0 5 10 15 20 25 30 35
Citation term weight
MAP
P@5
R-prec
bpref


KL score
0.8
0.6
0.4
0.2
0
0 5 10 15 20 25 30 35
Citation term weight
MAP
P@5
R-prec
bpref


KL FB score
0.8
0.6
0.4
0.2
0
0 5 10 15 20 25 30 35
Citation term weight
MAP
P@5
R-prec
bpref


Okapi score
0.8
0.6
0.4
0.2
0
0 5 10 15 20 25 30 35
Citation term weight
MAP
P@5
R-prec
bpref


Cosine score
0.8
0.6
0.4
0.2
0
0 5 10 15 20 25 30 35
Citation term weight
MAP
P@5
R-prec
bpref