How to Find Better Index Terms Through Citations

Anna Ritchie
University of Cambridge
Computer Laboratory
15 J J Thompson Avenue
Cambridge, CB3 0FD, U.K.
ar283@cl.cam.ac.uk

Simone Teufel
University of Cambridge
Computer Laboratory
15 J J Thompson Avenue
Cambridge, CB3 0FD, U.K.
sht25@cl.cam.ac.uk

Stephen Robertson
Microsoft Research Ltd
Roger Needham House
7 J J Thomson Avenue
Cambridge, CB3 0FB, U.K.
ser@microsoft.com

Abstract

We consider the question of how information from the textual context of citations in scientific papers could improve indexing of the cited papers. We first present examples which show that the context should in principle provide better and new index terms. We then discuss linguistic phenomena around citations and which type of processing would improve the automatic determination of the right context. We present a case study, studying the effect of combining the existing index terms of a paper with additional terms from papers citing that paper in our corpus. Finally, we discuss the need for experimentation for the practical validation of our claim.

1 Introduction

Information Retrieval (IR) is an established field and, today, the 'conventional' IR task is embodied by web searching. IR is mostly *term-based*, relying on the words within documents to describe them and, thence, try to determine which documents are relevant to a given user query. There are theoretically motivated and experimentally validated techniques that have become standard in the field. An example is the Okapi model; a probabilistic function for term weighting and document ranking (Spärck Jones, Walker & Robertson 2000). IR techniques using such statistical models almost always outperform more linguistically based ones. So, as statistical models are developed and refined, it begs the question 'Can Computational Linguistics improve Information Retrieval?'

Our particular research involves IR on scientific papers. There are definite parallels between the web and scientific literature, such as hyperlinks between webpages alongside citation links between papers. However, there are also fundamental differences, like the greater variability of webpages and the independent quality control of academic texts through the peer review process. The analogy between hyperlinks and citations itself is not perfect: whereas the number of hyperlinks varies greatly from webpage to webpage, the number of citations in papers is more constrained, due to the combination of strict page limits, the need to cite to show awareness of other work and the need to conserve space by including only the most relevant citations. Thus, while some aspects of web-based techniques will carry across to the current research domain, others will probably not. We are interested in investigating which lessons learned from web IR can successfully be applied to this slightly different domain.

2 Index Terms Through Link Structure

We aim to improve automatic indexing of scientific papers by finding additional index terms outside of the documents themselves. In particular, we believe that good index terms can be found by following the link structure between documents.

2.1 Hyperlinks

There is a wealth of literature on exploiting link structure between web documents for IR, including the 'sharing' of index terms between hyperlinked pages. Bharat & Mihaila (2001), for instance, propagate title and header terms to the pointed-to page, while Marchiori (1997) recursively augments the textual content of a page with *all* the text of the pages it points to.

Research has particularly concentrated on anchor text as a good place to find index terms, i.e., the text enclosed in the $\langle$a$\rangle$ tags of the HTML document. It is a well-documented problem that webpages are often poorly self-descriptive (e.g., Brin & Page 1998, Kleinberg 1999). For instance, www.google.com does not contain the phrase *search engine*. Anchor text, on the other hand, is often a higher-level description of the pointed-to page. Davison (2000) provides a good discussion of just how well anchor text does this and provides experimental results to back this claim. Thus, beginning with McBryan (1994), there is a trend of propagating anchor text along its hyperlink to associate it with the linked page, as well as that in which it is found. Google, for example, includes anchor text as index terms for the linked page (Brin & Page 1998).

Extending beyond anchor text, Chakrabarti et al. (1998) look for topic terms in a window of text around hyperlinks and weight that link accordingly, in the framework of a link structure algorithm, HITS (Kleinberg 1999).

2.2 Citations

The anchor text phenomenon is also observed with citations: they are introduced purposefully alongside some descriptive reference to the cited document. Thus, this text should contain good index terms for the cited document. In the following sections, we motivate the use of reference terms as index terms for cited documents, firstly, with some citation examples and, secondly, by discussing previous work.

Examples: Reference Terms as Index Terms

Figure 1 shows some citations that exemplify why reference terms should be good index terms for the cited document. (1) is an example of a citation with intuitively good index terms (those underlined) for the cited paper around it; a searcher looking for papers about a *learning system*, particularly one that uses *theory refinement* and/or one that learns *non-recursive NP and VP structures* might be interested in the paper, as might those searching for information about *ALLiS*.

The fact that an author has chosen those particular terms in referring to the paper means that they reflect what that author feels is important about the paper. It is reasonable, then, that other researchers interested in the same things would find the cited paper useful and could plausibly use such terms as query terms. It is true that the cited paper may well contain these terms, and they may even be important, prominent terms, but this is not necessarily the case. There are numerous situations in which the terms in the document are not the best indicators of what is important in it. Firstly, what is important in a paper in terms of what it is known and cited for is not always the same as what is important in it in terms of subject matter or focus. Secondly, what are considered to be the important contributions of a paper may change over time. Thirdly, the terminology used to describe the important contributions may be different from that used in the paper or may change over time.

(2) exemplifies this special case, where a paper is referred to using terms that are not in the paper itself: the cited paper is the standard reference for the HITS algorithm yet the name HITS was only attributed to the algorithm after the paper was written and it doesn't contain the term at all[1].

The last two examples show how citing authors can provide higher level descriptions of the cited paper, e.g., *good overview* and *comparison*. These meta-descriptors are less likely to appear in the papers themselves as prominent terms yet, again, could plausibly be used as query terms for a searcher.

Reference Directed Indexing

These examples (and many more) suggest that text used in reference to papers can provide useful index terms, just as anchor text does for webpages. Bradshaw & Hammond (2002) even go so far as to argue that reference is more valuable as a source of index terms than the document's own content. Bradshaw's theory is that, when citing, authors describe a document in terms similar to a searcher's query for the information it contains.

However, there *is* no anchor text, per se, in papers, i.e., there are no HTML tags to delimit the text associated with a citation, unlike in webpages. The question is raised, therefore, of what is the anchor text equivalent for formal citations. Bradshaw (2003) extracts NPs from a fixed window of around one hundred words around the citation and uses these as the basis of his *Reference-Directed Indexing* (RDI).

Bradshaw evaluates RDI by, first, indexing documents provided by Citeseer (Lawrence, Bollacker & Giles 1999). A set of 32 queries was created by randomly selecting keyword phrases from

[1] There is a poetic irony in this: Kleinberg's paper notes the analagous problem of poorly self-descriptive webpages.

(1) *ALLiS (Architecture for Learning Linguistic Structures) is a learning system which uses theory refinement in order to learn non-recursive NP and VP structures (Dejean, 2000).*

(2) Such estimation is simplified from HITS algorithm (Kleinberg, 1998).

(3) As two examples, (Rabiner, 1989) and (Charniak et al., 1993) give good overviews of the techniques and equations used for Markov models and part-of-speech tagging, but they are not very explicit in the details that are needed for their application.

(4) For a comparison to other taggers, the reader is referred to (Zavrel and Daelemans, 1999).

Figure 1: Citations Motivating Reference Index Terms

24 documents in the collection with an author-written keywords section. Document relevance was determined by judging whether it addressed the same topic as the topic in the query source paper that is identified by the query keywords. Thus, the performance of RDI was compared to that of a standard vector-space model implementation (TF*IDF term weighting and cosine similarity retrieval), with RDI achieving better precision at top 10 documents (0.484 compared to 0.318, statistically significant at 99.5% confidence).

Citing Statements

In a considerably earlier study, closer to our own project, 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 carried out 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 select the parts of the citing paper that conveyed information about the cited paper. For instance, the citing sentence, S, was always selected. If S contained a *connector* (a keyword, e.g., this, similarly, former) in its first twelve words, its predecessor, S_{-1}, was also selected etc. The majority of rules selected sentences from the text; others selected titles and words from tables, figures and captions.

The selected statements (minus stop words) were added to an existing representation for the cited documents, comprising human index terms and title 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%[2] (O'Connor 1983).

O'Connor concludes 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, which is a non-trivial problem in itself. In all sentence-based cases, sentences were either selected in their entirety or not at all and O'Connor notes this as a source of falsely assigned terms.

3 Complex Citation Contexts

There is evidence, therefore, that good index terms for scholarly documents can be found in the documents that cite them. Identifying which terms around a citation really refer to it, however, is non-trivial. In this section, we discuss some examples of citations where this is the case and propose potential ways in which computational linguistics techniques may be useful in more accurately locating those reference terms. We take as our theoretical baseline all terms in a fixed window around a citation.

3.1 Examples: Finding Reference Terms

The first two examples in Figure 2 illustrate how the amount of text that refers to a citation can vary. Sometimes, only two or three terms will refer to a citation, as is often the case in enumerations such as (5). On the other hand, (6) shows a citation where much of the following section refers to the cited work. When a paper is heavily based on previous work, for example, extensive text may be afforded to describing that work in detail. Thus, this context could contribute dozens of legitimate index terms. A fixed size window around a citation

[2] O'Connor attributes this to a lower average number of citing papers in the biomedical domain.

(5) *Similar advances have been made in machine translation (Frederking and Nirenburg, 1994), speech recognition (Fiscus, 1997) and named entity recognition (Borthwick et al., 1998).*

(6) Brown et al. (1993) proposed a series of statistical models of the translation process. IBM translation models try to model the translation probability ... which describes the relationship between a source language sentence ... and a target language sentence ... . In statistical alignment models ... a 'hidden' alignment ... is introduced, which describes a mapping from a target position ... to a source position ... . The relationship between the translation model and the alignment model is given by: ...

(7) The results of disambiguation strategies reported for pseudo-words and the like are consistently above 95% overall accuracy, far higher than those reported for disambiguating three or more senses of polysemous words (Wilks et al. 1993; Leacock, Towell, and Voorhees 1993).

(8) This paper concentrates on the use of zero, pronominal, and nominal anaphora in Chinese generated text. We are not concerned with lexical anaphora (Tutin and Kittredge 1992) where the anaphor and its antecedent share meaning components, while the anaphor belongs to an open lexical class.

(9) Previous work on the generation of referring expressions focused on producing minimal distinguishing descriptions (Dale and Haddock 1991; Dale 1992; Reiter and Dale 1992) or descriptions customized for different levels of hearers (Reiter 1990). Since we are not concerned with the generation of descriptions for different levels of users, we look only at the former group of work, which aims at generating descriptions for a subsequent reference to distinguish it from the set of entities with which it might be confused.

(10) Ferro et al. (1999) and Buchholz et al. (1999) both describe learning systems to find GRs. The former (TR) uses transformation-based error-driven learning (Brill and Resnik, 1994) and the latter (MB) uses memory-based learning (Daelemans et al., 1999).

Figure 2: Citations Motivating Computational Linguistics

would not capture all the terms referring to it and only those.

In list examples such as (5), where multiple citations are in close proximity, almost any window size would result in overlapping windows and in terms being attributed to the wrong citation(s), as well as the right one. In such examples, the presence of other citations indicates a change in reference term 'ownership'. The same is often true of sentence boundaries, as they often signal a change in topic. Citations frequently occur at the start of sentences, as in (6), where a different approach is introduced. Similarly, a citation at the end of a sentence, as in (7), often indicates the completion of the current topic. In both cases, the sentence boundary (c.f. topic change) is also the boundary of the reference text. The same arguments increasingly apply to paragraph and section boundaries.

(8) is another example where the reference text does not extend beyond the citation sentence, though the citation is not at a sentence boundary. Instead, the topic contrast is indicated by a linguictic cue, i.e., the negation in *We are not*. This illustrates another phenomenon of citations: in contrasting their work with others', researchers often explicitly state what their paper is *not* about. Intuitively, not only are these terms better descriptors of the cited rather than citing paper, they might even raise the question of whether one should go as far as *excluding* selected terms during indexing of the citing paper. We are not advocating this here, though, and note that, in practice, such terms would not have much impact on the document: we would expect them to have low term frequencies in comparison to the important terms in that document and in comparison to their frequencies in other documents where they *are* important.

(9) is another example of this negation effect (*We are not concerned with...*). Along with (10), it also shows how complex the mapping between reference terms and citations can be. Firstly, reference terms may belong to more than one cita-

tion. For instance, in (10), *describe learning systems to find GRs* refers to both *Ferro et al. (1999)* and *Buchholz et al. (1999)*. Here, the presence of a second citation does not end the domain of the first's reference text, indicated by the use of *both* and the conjunction between the citations. Similarly, *transformation-based error-driven learning* also refers to two citations but, in this case, they are on opposite sides of the reference text, i.e., *Ferro et al. (1999)* and *(Brill and Resnik, 1994)*. Moreover, there is an intervening citation that it does not refer to, i.e., *Buchholz et al. (1999)*. The same is true of *memory-based learning*.

4 Case Study

In this section, we study the effect of adding citation index terms to one document: *The Mathematics of Statistical Machine Translation: Parameter Estimation* from the Computational Linguistics journal[3]. Our experimental setting is a corpus of ∼9000 papers in the ACL Anthology[4], a digital archive of computational linguistics research papers. We found 24 citations to the paper in 10 other Anthology papers (that we knew to have citations to this paper through an unrelated study). As a simulation of ideal processing, we then manually extracted the terms from those around those citations that specifically referred to the paper, henceforth *ideal reference terms*. Next, we extracted all terms from a fixed window of ∼50 terms on either side (equivalent to Bradshaw (2003)'s window size), henceforth *fixed reference terms*. Finally, we calculated various term statistics, including IDF values across the corpus. All terms were decapitalized. We now attempt to draw a 'term profile' of the document, both before and after those reference terms are added to the document, and discuss the implications for IR.

4.1 Index Term Analysis

Table 1 gives the top twenty ideal reference terms ranked by their TF*IDF values in the original document. Note that we observe the effects on the relative rankings of the ideal reference terms only, since it is these hand-picked terms that we consider to be important descriptors for the document and whose statistics will be most affected by the inclusion of reference terms. To give an indication of their importance relative to other terms in the document, however, the second column in Table 1 gives the absolute rankings of these terms in the original document. These numbers confirm that our ideal reference terms are, in fact, relatively important in the document; indeed, the top five terms in the document are all ideal reference terms. Further down the ranking, the ideal reference terms become more 'diluted' with terms not picked from our 24 citations. An inspection revealed that many of these terms were French words from example translations, since the paper deals with machine translation between English and French. Thus, they were bad index terms, for our purposes.

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 1: Ideal Reference Term Ranking by TF*IDF

Hence, we observed the effect of adding, first, the ideal reference terms then, separately, the fixed reference terms to the document, summarized in Tables 2 to 5. Tables 2 and 3 show the terms with the largest differences in positions as a result of adding the ideal and fixed reference terms respectively.

For instance, *ibm*'s TF*IDF value more than doubled. The term *ibm* appears only six times in the document (and not even from the main text but from authors' institutions and one bibliography item) yet one of its major contributions is the machine translation models it introduced, now standardly referred to as 'the IBM models'. Con-

[3] http://www.aclweb.org/anthology/J93-2003.pdf

[4] http://www.aclweb.org/anthology/

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: Term Ranking Changes (Ideal)

Term	TF*IDF		Ideal
	Δ	Doc+fixed	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

Table 3: Term Ranking Changes (Fixed)

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 4: New Non-zero TF*IDF Terms (Ideal)

sequently, ‘IBM’ was contained in many citation contexts in citing papers, leading to an ideal reference term frequency of 11 for *ibm*. As a result, *ibm* is boosted eight places to rank 20. This exemplifies how reference terms can better describe a document, in terms of what searchers might plausibly look for (c.f. Example 2).

There were twenty terms that do not occur in the document itself but are nevertheless used by citing authors to describe it, shown in Tables 4 and 5. Many of these have high IDF values, indicating their distinctiveness in the corpus, e.g., *decoders* (6.41), *corruption* (6.02) and *noisy-channel* (5.75). This, combined with the fact that citing authors use these terms in describing the paper, means that these terms are intuitively high quality descriptors of the paper. Without the reference index terms, however, the paper would score zero for these terms as query terms.

Many more fixed reference terms were found per citation than ideal ones. This can introduce noise. In general, the TF*IDF values of ideal reference terms can only be further boosted by including more terms and a comparison of Tables 2 with 3 (or 4 with 5) shows that this is sometimes the case, e.g, *ibm* occurred a further eleven times in the fixed reference terms, doubling its increase in TF*IDF. However, instances of those terms that only occurred in the fixed reference terms did not, in fact, refer to the citation of the paper, by definition of the ideal reference terms. For instance, one such extra occurrence of *ibm* is from a sentence following the citation that describes the exact model used in the current work:

(11) *According to the IBM models (Brown et al., 1993), the statistical word alignment model can be generally represented as in Equation (1) ... In this paper, we use a simplified IBM model 4 (Al-Onaizan et al., 1999), which ...*

Here, the second occurrence refers to *(Al-Onaizan et al., 1999)* but, by its proximity to the citation to our example paper *(Brown et al., 1993)*, is picked up by the fixed window. Since the term was arguably not directly intended to describe our paper, then, a different term might equally have been used; one that was inappropriate as an index term. Table 6 lists the fixed reference terms that were not also in the ideal reference terms; almost 400 in total. The vast majority of these occur very infrequently which suggests that they should not greatly affect the term profile of the document. However, the argument for adding *good*, high IDF reference terms that are not in the document itself

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

Table 5: New Non-zero TF*IDF Terms (Fixed)

conversely applies to adding *bad* ones: an ‘incorrect’ reference term added to the document will have its TF*IDF pushed off the zero mark, giving it the potential to score against inappropriate query terms. If such a term is distinctive (i.e., has a high IDF), the effect may be significant. The term *giza*, for example, has an IDF of 6.34 and is the name of a particular tool that is not mentioned in our example paper. However, since the tool is used to train IBM models, the two papers in the example above are often cited by the same papers and in close proximity. This increases the chances of such terms being picked up as reference terms for the wrong citation by a fixed window, heightening the adverse effect on its term profile.

5 Discussion and Conclusions

It is not too hard to find examples of citations that show a fixed window size is suboptimal for finding terms used in reference to cited papers. In extracting the ideal reference terms from only 24 citations for our case study, we saw just how difficult it is to decide which terms refer to which citations. We, the authors, came across examples where it was ambiguous how many citations certain terms referred to, ones where knowledge of the cited papers was required to interpret the scope of the citation and ones where we simply did not agree. This is a highly complex indexing task; one which humans have difficulty with, one for which we expect low human agreement and, therefore, the type that computational linguistics struggles to achieve high performance on. We agree with O’Connor (1982) that it is hard. We make no claims that computational linguistics will provide a full solution.

Nevertheless, our examples suggest that even simple computational linguistics techniques should help to more accurately locate reference terms. While it may be impossible to automatically pick out each specific piece of text that does refer to a given citation, there is much scope for improvement over a fixed window. The examples in Section 2 suggest that altering the size of the window that is applied would be a good first step. Some form of text segmentation, whether it be full-blown discourse analysis or simple sentence boundary detection, may be useful in determining where the extent of the reference text is.

While the case study presented here highlights several interesting effects of using terms from around citations as additional index terms for the cited paper, it cannot answer questions about how successful a practical method based on these observations would be, over a using simple fixed window, for example. In order for any real improvement in IR, the term profile of a document would have to be significantly altered by the reference terms. Enough terms, in particular repeated terms, would have to be successfully found via citations for such a quantitative improvement. It is not clear that computational linguistic techniques will improve over the statistical effects of redundant data.

We are thus in the last stages of setting up a larger experiment that will shed more light on this question. The experimental setup requires data where there are a significant number of citations to a number of test documents and a significant number of reference set terms. We have recently presented a test collection of scientific research papers (Ritchie, Teufel & Robertson 2006), which we intend to use for this experiment.

References

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

Bradshaw, S. (2003), Reference directed indexing: Redeeming relevance for subject search in citation indexes., *in* ‘ECDL’, pp. 499–510.

Bradshaw, S. & Hammond, K. (2002), Automatically indexing documents: Content vs. reference, *in* ‘Intelligent User Interfaces’.

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

Table 6: Term Frequencies of 'Noisy' Reference Index Terms

Brin, S. & Page, L. (1998), 'The anatomy of a large-scale hypertextual Web search engine', *Computer Networks and ISDN Systems* **30**(1–7), 107–117.

Chakrabarti, S., Dom, B., Gibson, D., Kleinberg, J., Raghavan, P. & Rajagopalan, S. (1998), Automatic resource list compilation by analyzing hyperlink structure and associated text, *in* 'Seventh International World Wide Web Conference'.

Davison, B. D. (2000), Topical locality in the web, *in* 'Research and Development in Information Retrieval (SIGIR)', pp. 272–279.

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

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

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

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

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.

Ritchie, A., Teufel, S. & Robertson, S. (2006), Creating a test collection for citation-based IR experiments, *in* 'HLT-NAACL'.

Spärck Jones, K., Walker, S. & Robertson, S. E. (2000), 'A probabilistic model of information retrieval: development and comparative experiments - parts 1 & 2.', *Information Processing and Management* **36**(6), 779–840.