Optimal citation context window sizes for biomedical retrieval

Nielsen, Boris Lykke; Skau, Stefan Lavlund; Meier, Florian; Larsen, Birger

Published in:
CEUR Workshop Proceedings

Publication date:
2019

Document Version
Publisher's PDF, also known as Version of record

Optimal Citation Context Window Sizes for Biomedical Retrieval

Boris Lykke Nielsen, Stefan Lavlund Skau, Florian Meier, and Birger Larsen

Science, Policy and Information Studies
Department of Communication and Psychology
Aalborg University, Copenhagen, Denmark
{fmeier,birger}@hum.aau.dk

Abstract. We investigate the TREC-CDS 2016 test collections as a new resource for citation context and citation-based IR experiments. The collection contains more than 1.25 million biomedical full-text articles in XML. We find that a citation index can easily be extracted, and citation contexts easily be identified. We conduct initial experiments to determine the optimal citation context window size in this domain and collection. Surprisingly We find that quite long citation contexts of more than 250 word yield the best performance when combined linearly with the full-text and with moderate weight on the citation contexts.

Keywords: Citation contexts for IR · TREC-CDS 2016 · Citation Context Windows

1 Introduction

In this work, we investigate the feasibility of extracting citation contexts from citing articles and using them in the retrieval of scientific documents. Adding citation contexts of a document to its full-text may help the retrieval process by providing additional relevant keywords for indexing. Keywords from citation contexts may be valuable as they provide a different perspective on the cited text — that of the authors citing and using the cited document.

Prior research on this idea indicates that citation contexts can indeed improve retrieval performance, e.g. [14,1,16,3]. Most previous work, however, has been carried out on small collections of documents of no more than a few thousand documents or in specialized domains. In the present work, we take advantage of the increased availability of Open Access publications in full-text to extract and study the usefulness of citation contexts for scientific retrieval in the hitherto largest publicly available test collection that supports this type of retrieval. Specifically, we work with 1.25 million documents from the biomedical domain taken from the Open Access Subset of the PubMed Central collection. These documents were included in the 2016 Text REtrieval Conference Clinical Decision Support track (TREC-CDS) which produced a test collection that in addition to the documents also includes information needs and associated relevance assessments by medical professionals [17]. This allows us to carry out

experiments where we explore the feasibility of extracting citation contexts from such a vast collection and their potential for improving retrieval performance in this domain. In particular as an initial effort we investigate the following research question: What is the optimal citation context window size with respect to improving retrieval performance?

The paper is structured as follows: Section 2 discusses related work. Section 3 presents the methods we used including analysis of the TREC-CDS 2016 test collection, and details on the extraction of citation contexts. Section 4 describes our experimental setup and our findings. Section 5 presents discussion and conclusions.

2 Related work

In establishing the Science Citation Index in 1964 Eugene Garfield created a retrieval system solidly based on the idea that citations form explicit links between papers that have particular points in common and can be used to search the scientific literature [7, p.1]. A researcher would rely on the author's judgement to include references to other publications with shared subjects or topics and thus, through the network established from bibliographic references, identify other citing or cited papers with similar or new topics or subjects in common [7, p.2]. Continued as Clarivate's Web of Science, Garfield's citation indexes is still based on the links between papers, but ignores the nature and meaning of these links. With the increasing availability of scientific literature in full-text and as open source, we can now begin to investigate the nature of the links by studying the text where a given paper is mentioned. Existing research has demonstrated that the text surrounding citations often contains descriptions of the cited paper, the reason or function of the citation, or the disposition towards the cited paper [10, p.201]. As such, it may be possible to glimpse what the cited paper is about, or how the author has used the paper, by examining the surrounding associated text of a specific citation. In other words, by examining the textual content of a citation, it is potentially possible to identify how and why a citation is made.

Analysis of Citation Contexts The purpose of citation context analysis, first proposed by Small[18], is in general to examine the contextual relationship between the citing and cited papers. White [23] reviews work in the area and notes three lines of research: (1) *Classifying citations*, which "are attempts to understand what people are doing when they cite" and involve citation classification schemes of which he identifies and compares more than 20, (2) *Content analysis of citation contexts*, in which words occurring in citation contexts are used to describe the cited paper as basis for analysis, and (3) studies of *Citer motivations*, which examines the deeper reason for "*why* authors make references". In the present paper we do content analysis of citation contexts on a large scale.

Citation context analysis has been studied from a number of perspectives, including citation summarisation [13,5,15,2], creation of personalised citation

recommendations [11], discovery of new knowledge [20] or to manually or automatically classify the motivations and functions that lies behind citations and measure their impact [9,22].

Analysis of Citation Contexts and IR Performance Of particular interest to the present work is attempts to enhance and improve retrieval of scientific documents. An early example is O'Connor who proposed to extract additional indexing terms from citation contexts that cite a given paper [14]. Bradshaw follows a similar approach, but argues that terms from citation contexts can suffice as document representation [1]. Both authors demonstrate that terms from citation contexts can indeed improve retrieval performance — in particular precision-based measures. More recent works that have used citation contexts to improve retrieval of literature and have further motivated our project, are the works by Ritchie [16] and Dabrowska and Larsen [3]. Ritchie conducted retrieval experiments on a small test collection of 9800 full-text documents within the scientific area of computational linguistics [16]. Their collection contained approximately 20.000 citations pointing to about 3200 documents within the collection. In her work, Ritchie defined citation contexts, similar to Bradshaw, within fixed windows, but pursued more variations of windows using both sentences and words (50, 75, 100 on each side of the citation), to be able to compare the effectiveness of the window sizes relative to one another.

Following a similar approach to Ritchie but at a larger scale, Dabrowska and Larsen performed retrieval experiments on a test collection of over 430.000 full-text papers from a subset of the iSearch (Integrated Search) document collection [3]. This collection contains approximately 3.7 million citations, with about 260.000 unique documents being cited by other documents within the collection [12]. They found that retrieval, with their larger collection of physics papers, was improved with the addition of citation contexts as index terms to the full-text documents [3]. More specifically, they found improvements with fixed windows of words (25, 50, 75 & 100 on each side) with the best results having a moderate weight (of around 25%) on the citation contexts relative to the full-text.

3 Method

3.1 Data-set

For our experiments we use the *TREC 2016 Clinical Decision Support Track*[1] data-set [17]. Like most data-sets released by TREC, it consists of (i) a collection of documents (ii) topics or user search tasks (iii) relevance assessments on which documents of the collection are relevant for the topics. The document collection is made up of 1.25 million full-text biomedical articles representing a snapshot of the *Open Access Subset* of *PubMed Central* (PMC)[2]. PMC launched in 2000

[1] http://www.trec-cds.org/2016.html
[2] https://www.ncbi.nlm.nih.gov/pmc/

and is a free archive for full-text biomedical and life sciences journal articles. It contains at present more than 5 million full-text articles, most of which are also included in as bibliographical references in the 25+ million records of PubMed. Each article in the collection is represented as an NXML file and identified by an official PMCID.

The topics of the track simulate actual information needs of physicians and are divided in three different types representing the most common generic clinical questions [6]. These types are: (i) **Diagnosis**, (ii) **Treatment** and (iii) **Test**. A search task consists of (i) An admission note, (ii) A case description based on the note, (iii) A summary of the case. These summaries were often written as shortened versions of the case description. The search queries used in the experiments of this study were based on the summary sections (iii), and the summaries were not edited in any way when before being inserted into *Indri* as search queries. The relevance assessments were conducted after the submission of retrieval runs by participants of the track [17]. The relevance assessment was done by pooling results from the submitted retrieval runs. The pooled documents were then assessed as *Definitely relevant (2)*, *Possibly relevant (1)*, and *Not relevant (0)* [17]. A rating of *Possibly relevant* was given to documents that were not themselves relevant to the topic but could prove relevant in the context of a broader literature review [17]. The released relevance assessments (so-called QRELs) contain 28,349 unique documents, 5,461 of which are *Definitely* or *Possibly relevant*.

3.2 Extracting Citation Contexts

Citation contexts are defined as the textual passages or sentences surrounding or containing the citations [19]. In this work we utilize the cross-reference tags, to determine the position and target reference for each in-text reference and extract citation contexts. Although the position of the citation in the text is known, identifying the start and end, i.e. the optimal context length is a difficult and complex problem [9]. Several researchers have used windows of fixed sizes starting with O'Connor and Bradshaw. However, by using fixed windows to define the contexts, it is possible that the context does not adequately characterise the relationship to the referred citation. The context might exclude words or sentences that implicitly refers to the citation, or include words or sentences that do the exact opposite. In other words, the defined citation context should only contain the text that describes the cited paper. This has been attempted by considering the linguistic features of the text to define contexts within windows that contains the full scope of descriptive text to identify the optimal context window size around the citation [16,8].

We downloaded the TREC-CDS 2016 collection and first investigated the raw XML files to determine if (i) internal citations between the documents in the data-set can be readily identified, and to answer the question: (ii) how feasible is it to identify and extract citation contexts?

The documents in the TREC-CDS 2016 test collection are encoded in the Journal Archiving and Interchange Tag Set (JATS)[3]. All full-text documents in the test collection use the XML file type and use the tags defined in the JATS DTD. The JATS include special tags for bibliographic references (`<xref>`) which are wrapped around cross-references in the full-text document. The conducted experiments utilised the cross-reference tags to determine the position and target reference for each in-text reference. Additionally, the JATS include special tags for the reference list to ease parsing of the list. This list was used to determine the target document for each in-text reference. The documents in the test collection have two unique and different identifiers that we could use to map documents to each other: (i) **PubMed Central Identifier (PMCID)** This identifier is assigned to all documents included in the PMC. As the test collection is a snapshot of a subset of the PMC, all documents will have this identifier. (ii) **PubMed Identifier (PMID)** This identifier is assigned to any record also included in PubMed. This identifier (and not the PMCID) has also been added by PMC staff in the JATS reference lists when pointing to documents included in PubMed.

As each document has a PMCID as well as a PMID (if it is in PubMed) in its header, we were able to match these two IDs as a basis for extracting a citation network and citation contexts. It is important to note that because not all documents in the PMC are also present in PubMed, not all documents in the test collection have a PMID. 87.622 (7%) of the documents did not have a PMID.

Table 1 gives a summary of the statistics of the extracted data. Of the 1.25 million documents just over a million have references (87.3%). 58,845 of the documents without references are abstract-only documents without full-text, and the remaining may be publication types that do not contain references (e.g. editorials, news items, etc.) The 1+ million documents with references contain more than 43 million references (40 references on average per article). Of the 1,255,260 documents in the collection, 567,650 documents (45.2%) received at least one citation from within the collection. 370,426 of these were cited between 2 and 100 times — see Table 1 for other ranges. More than 60 million citation contexts were identified (the same document can be mentioned more than once in the full-text in one document). 46 million of these (76.6%) have a target PMID pointing to a PubMed document, and 4.8 million (8%) could be matched to a PMCID in the TREC-CDS 2016 collection. These 4,833,813 citation contexts point to the 567,650 cited documents and form the core data-set used in our experiments. This means that each cited document has 8.5 linked citation contexts on average. A total of 37,707 documents were assessed for relevance in relation to the 30 topics. Some were retrieved and assessed for multiple topics. The number of unique PMCIDs in the QRELs is 28,349. Of these, 10,132 (36.1%) were cited and had at least one associated citation context. However, only 2,019 of these were assessed as *Definitely relevant* or *Possibly relevant*. This means that only 37% of the relevant documents also had citation contexts added.

[3] https://jats.nlm.nih.gov/index.html

Table 1. Summary statistics on the TREC-CDS 2016 test collection and extracted citation contexts.

Statistic	# of docs/refs/contexts
Documents scanned	1,255,260
Documents with references	1,096,062
Abstract-only documents	58,845
References total count	43,840,755
Cited once or more inside test collection	567,650
Cited once	196,232
Cited between 2 and 100 times	370,426
Cited between 101 and 1000 times	973
Cited more than 1001 times	18
Contexts total count	60,105,922
Contexts with target PMID	46,027,299
Usable contexts (i.e. with a PMCID)	4,833,813
Documents with appended contexts	567,650
Length of QREL	37,707
Unique PMCIDs in QREL	28,349
Unique and relevant documents in QREL	5,461
Documents with appended contexts in QREL	10,132
Relevant documents with contexts in QREL	2,019

4 Experiments

4.1 Experimental Setup

Citation Context lengths We limit ourselves to the simplistic, however, widely used approach of using fixed window sizes surrounding the citation. This is done by considering the citation as a starting point and then extend the context to include text around the citation with some determined fixed or variable distance on each side of the citation. Citation context text can be variable and range from just a few characters, over phrase and clauses to sentences and entire paragraphs. The task of identifying the optimal context length of citations is difficult and complex, as the citing behaviour, characteristics and processes of citations is different within papers and sections and across authors, scientific discourses, fields and domains [4]. However, one can argue that the best context length is the one that yields the best results for the given purposes. In the present work we investigate two simple approaches: (i) extracting a number of sentences before and/or after the sentence in which the in-text reference occurs, and (ii) extracting a number of words before and/or after the in-text reference. We did experiments with 1-5 sentences before the in-text reference, and/or 1 sentence after the reference and with 50-300 words on the left and/or right of the in-text reference (See columns one and two of Table 2 for an overview. 300_25 means that we used 300 words, with 25% = 75 words on the right). We include more text before the in-text reference as these often occur at the end of a sentence,

and as we expect that most of the text commenting on that reference occurs before it. Ritchie [16] and Dabrowska and Larsen [3] used up to 100 words. The maximum of 300 words correspond to almost a page of full-text in the format of the present paper, and we deemed it to be well beyond the upper-bound for what could be beneficial for retrieval.

Retrieval experiments As we are working with a test collection, our experiments are solidly within the Cranfield tradition. We use the *Indri* IR system to conduct the retrieval experiments, with Language Modeling and Dirichlet smoothing [21]. Initial experimentation in relation to constructing a baseline investigated 24 values of the μ tuning parameter of the Dirichlet smoothing between 1 and 45,000. Values in the range 20,000 - 35,000 provided the better results with this collection so we choose to tune μ in this range: 20000, 22000, 24000, 27000, 30000, 33000, 36000, 39000. The baseline was tuned separately for P@10, MAP and NDCG - the resulting baseline values can be seen in Table 2 [4].

To integrate citation contexts seamlessly into retrieval we took advantage of two features in *Indri*. First, we placed the full-text of the original article in a separate field and added the different versions of extracted citation contexts in fields of their own. Second, we used the Indri query language to create a linear combination of the two. We used the #weight operator to assign weights to the full-text and citation context respectively in each query, with the weights adding up to 1. The following is an example of the query formatting (topic 1 from TREC-CDS 2016):

```
#weight (0.6 #combine [fulltext] (A 78 year old male presents with
frequent stools and melena) 0.4 #combine [250_25] (A 78 year old male
presents with frequent stools and melena))
```

where the same query string is matched against the full-text field with a weight of 0.6, and against one of the citation context fields with a weight of 0.4 (the #combine operator is standard operator for combining beliefs in *Indri*). In this way we can add the citation contexts as additional index terms to the cited document, and control their influence relative to the full-text. We tested weights of 20, 40, 60, 80 and 100% in the main experiment. Retrieval results were evaluated using *trec_eval*. For P@10 and MAP both *Definitely relevant* and *Possibly relevant* were counted as relevant — for NDCG *Definitely relevant* had a gain value of 2 and *Possibly relevant* a gain value of 1. P@10 represents a purely precision-oriented perspective on results whereas MAP and NDCG gives a perspective that balances precision and recall. It should be noted that retrieved documents that were not assessed were counted as non-relevant in all the measures.

[4] The smoothing parameter for the baseline run was 24,000 for MAP and 30,000 for MAP and NDCG.

4.2 Findings

The aim of the experiments was to learn how much context to include to reach optimal retrieval performance, and to determine how much weight to put on them relative to the full-text. Table 2 shows the overall results for P@10, MAP and NDCG. The table shows the best result for each run with the optimal smoothing parameter and the weight between full-text and citation context that performs the best. Figures 1 and 2 below demonstrate the effect of weighting and smoothing.

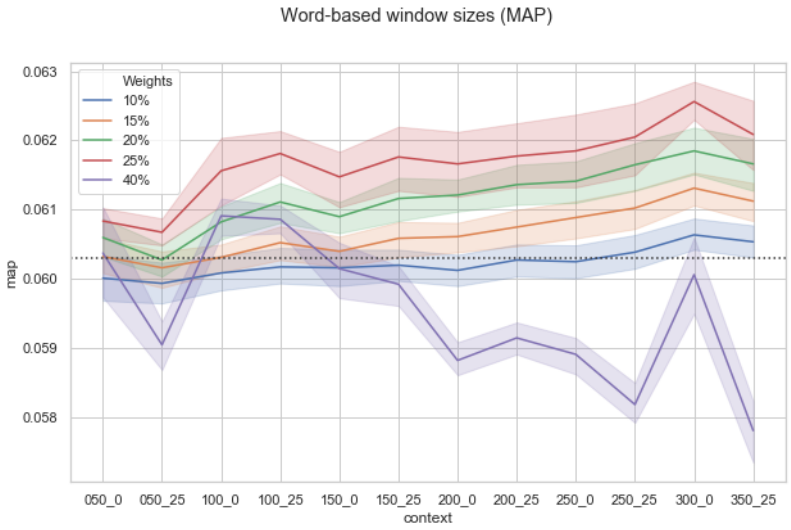

Fig. 1. Effect of the linear combination of full-text and citation contexts on MAP. Weights of 10, 15, 20, 25 and 40% on the citation contexts shown across the word-based citation context windows (see Table 2 for Run IDs.) Dotted line is the baseline.

From Table 2 we see that the citation context runs all outperform the baseline to some degree. The runs with more context added in general perform better. The best performing run for P@10 is the one with 250 words added (19% over the baseline), and for MAP and NDCG the run with 300 words added (4.6% and 2.8% over the baseline respectively). The best performing word-based runs outperform the sentence-based runs. An explanation may be that the sentence-based runs in general would be shorter in term of the number of words - with the longest of 5 sentences corresponding to 150 words or less. [5]

[5] We did not examine sentence length of the 4.8 million citation contexts, but other studies of academic biomedical text find around 25 words per sentence on average [24].

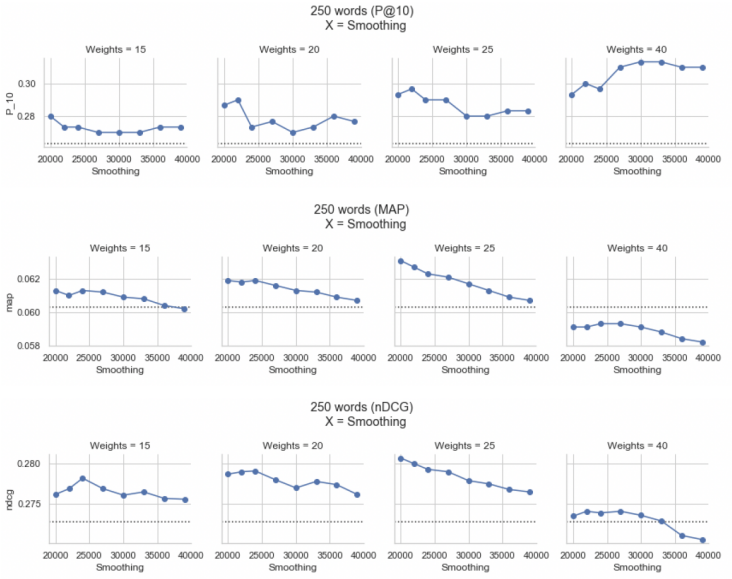

Fig. 2. Effect of smoothing

With regards to whether adding text to the right of the in-text reference is beneficial, a more complex picture emerges: for early precision (P@10) runs without right-hand text performs better in almost all cases — especially for the better performing runs with more words. For MAP and NDCG the actual differences are very small — for the runs with more context in the 100-250 word range performance is marginally better with right-hand text, but as more context gets added, right-hand context doesn't seem to have much influence, as can be seen for the MAP performance (4.6% over baseline) in the 250-300 word range.

Overall, all results point to the fact that more context added leads to better performance. Given this finding even larger context windows should be experimented with to ensure that an upper-bound has indeed been reached at 300 words, which was the maximum context length studied in the present experiment.

The results in Table 2 represent the best smoothing and weight combinations. With respect to the linear combination the top 8 runs in relation to P@10 all have 40% weight on citation contexts - for MAP and NDCG the top runs all have a 20% weight on the citation contexts (not shown in Table 2). Figure 1 demonstrates the effect of the linear combination of full-text and the added citation contexts for MAP for weights of 10, 15, 20, 25 and 40%. The shading of each line indicates the variation across the μ parameter tuning range, and the dashed line is the baseline. It can be seen that performance is very stable, and increases steadily as more weight is placed on the citation contexts from 10% up to 25%. This also holds across the different context length combinations with the

Table 2. Main retrieval results for optimal citation context window sizes. Average P@10, MAP and NDCG over 30 topics for sentence and word-based citation contexts. Scores for highest performing linear combination shown, with percentage increase over the baseline. Best runs for each measure is highlighted in blue.

	RunID	P@10	MAP	NDCG
baseline	-	**0.2633**	**0.0603**	**0.2736**
1 sentence	-	0.2767 (5.1%)	0.0611 (1.3%)	0.2766 (1.4%)
3 sentences	-	0.2933 (11.4%)	0.0612 (1.5%)	0.2766 (1.4%)
4 sentences	-	0.2933 (11.4%)	0.0616 (2.2%)	0.2780 (1.9%)
4 sentences, 1 right	-	0.2900 (10.1%)	0.0612 (1.5%)	0.2773 (1.6%)
6 sentences	-	0.2867 (8.9%)	0.0619 (2.7%)	0.2789 (2.2%)
50 words	050_0	0.2767 (5.1%)	0.0615 (2.0%)	0.2770 (1.2%)
50 words, 25% right	050_25	0.2933 (11.4%)	0.0612 (1.5%)	0.2767 (1.1%)
100 words	100_0	0.2933 (11.4%)	0.0625 (3.6%)	0.2781 (1.6%)
100 words, 25% right	100_25	0.2867 (8.9%)	0.0626 (3.8%)	0.2788 (1.9%)
150 words	150_0	0.3067 (16.5%)	0.0621 (2.9%)	0.2785 (1.8%)
150 words, 25% right	150_25	0.3000 (13.9%)	0.0625 (3.6%)	0.2790 (2.0%)
200 words	200_0	0.3067 (16.5%)	0.0624 (3.5%)	0.2798 (2.3%)
200 words, 25% right	200_25	0.3033 (15.2%)	0.0629 (4.3%)	0.2794 (2.1%)
250 words	250_0	0.3133 (19.0%)	0.0631 (4.6%)	0.2780 (2.6%)
250 words, 25% right	250_25	0.3100 (17.7%)	0.0631 (4.6%)	0.2803 (2.4%)
300 words	300_0	0.3033 (15.2%)	0.0631 (4.6%)	0.2811 (2.8%)
300 words, 25% right	300_25	0.3000 (13.9%)	0.0630 (4.5%)	0.2807 (2.6%)

top performance being reached at 300 words as discussed above, with scores that are consistently over the baseline except for the shortest citation context windows. At 40% however, performance overall drops below the baseline and shows great diversity across contexts lengths, demonstrating that too much weight on the citation contexts can hurt performance and lead to erratic behaviour. It can also be noted that variation across the smoothing parameter range (shading) is not prohibitively large with little overlap between each type of weighting.

Figure 2 further illustrates the effect of smoothing for contexts of 250 words (the best performing context length for P@10). Results are plotted for weights of 15, 20, 25 and 40% and for P@10, MAP and NDCG. A low or moderate level of variation across the tuning range is desirable as such an approach is less dependent on setting the tuning parameter correctly and can thus be considered more stable. It can be observed that performance is quite stable across the chosen smoothing range, with a bit more variation for P@10, which can be expected as it is a less stable measure. It can be clearly seen that a weight of 40% outperforms other weights for P@10 across the smoothing range. At a weight of 20% MAP and NDCG is very stable across the smoothing range. Higher performance can be achieved at 25% for MAP and NDCG but only at the lower smoothing values.

Finally, it should be noted that as expected, the citation distribution is skewed (1) - with 63% receiving no citations, the remaining 37% receiving more

than 1 - and 18 documents receiving more than 1000 citations. With the experimental setup used this means that some documents have no additional representation, and that a sizable proportion have a great deals of text added to their representation - in some cases tens of thousands of words. The effect of this on retrieval is at present unknown.

5 Discussion and Conclusion

Generating a citation index and extraction of citation contexts It was unproblematic to identify internal citations documents in the TREC-CDS 2016 collection due to explicit tagging in the XML and the added PMIDs in the reference lists. Not all documents have references, but more than 1 million do — providing a rich testbed for citation-based IR. Further, 45% of the documents in the collection received at least one internal citation. The resulting citation index from RQ1 and the XML and JATS formats made it possible to identify citing documents and to identify the location of in-text references. 4,8 million citation contexts could thus be extracted and linked to the 567,650 cited documents. It is worth noting that only 37% of the relevant documents had one or more citation contexts added - this reduces the impact that citation contexts can have on IR performance in the TREC-CDS collection. This underlines that fact that retrieval based on citation contexts is inherently dependent on documents being cited, and that a test collection which has been created with citation and citation context based runs in the pooling process is really needed to fully investigate the true potential of these approaches.

We limited ourselves to a simple definition of citation contexts extracting a number of words or sentences before and/or after each in-text reference. Tests of more advanced linguistically-based definitions would be interesting, but can be a challenge given the large number of contexts. Somewhat to our surprise the best performance was found among the longest citation windows of 250-300 words - both for precision and recall-oriented measures. As argued this a large amount of text (a full page) that almost certainly goes beyond where a given cited document is discussed. This may indicate that identifying the exact extent of the actual citation context may not be of prime importance - and on the other hand leads to the question of why so much text from citing documents is beneficial for retrieval, and if even larger windows will be beneficial?

Compared to previous research this is much longer citation context windows than previously tested. With regards to how much weight to put on the contexts our findings are in line with previous work of e.g. Ritchie (2009) and Dabrowska (2014). Best performance is achieved with moderate weight on the citation contexts of around 20% relative to the full-text - much more leads to decreased performance and erratic behaviour. As regards stability, results are quite stable across the smoothing range, indicating that this approach does not depend critically on getting the smoothing parameter right.

The present study mainly serves to introduce the TREC-CDS 2016 test collection as an attractive ressource for the BIR community and those interested

in citation context analysis - and to conduct initial tests of the feasibility of IR experiments using such features on this collection. Much more interesting work, where it is tested if the context of citations are useful for semantically categorizing a relationship and perhaps even an intention or an opinion between two publications, can be built on top of this, e.g. along the lines of Ritchie [16].

5.1 Acknowledgments

We wish to thank the organisers of TREC-CDS for creating a great resource, and the four anonymous reviewers for insightful comments.

References

1. Bradshaw, S.: Reference directed indexing: Redeeming relevance for subject search in citation indexes. In: Koch, T., Sølvberg, I.T. (eds.) Research and Advanced Technology for Digital Libraries. pp. 499–510. Springer, Berlin, Heidelberg (2003)
2. Cohan, A., Goharian, N.: Scientific document summarization via citation contextualization and scientific discourse. Int. J. Digit. Libr. **19**(2-3), 287–303 (2018)
3. Dabrowska, A., Larsen, B.: Exploiting citation contexts for physics retrieval. Proc. of BIR Workshop @ ECIR pp. 14–21 (2015)
4. Ding, Y., Zhang, G., Chambers, T., Song, M., Wang, X., Zhai, C.: Content-based citation analysis: The next generation of citation analysis. JASIST **65**(9), 1820–1833 (2014)
5. Elkiss, A., Shen, S., Fader, A., Erkan, G., States, D., Radev, D.: Blind men and elephants: What do citation summaries tell us about a research article? JASIST **59**(1), 51–62 (2008)
6. Ely, J.W., Osheroff, J.A., Gorman, P.N., Ebell, M.H., Chambliss, M.L., Pifer, E.A., Stavri, P.Z.: A taxonomy of generic clinical questions: classification study. BMJ **321**(7258), 429–432 (2000)
7. Garfield, E.: Citation indexing - Its Theory and Application in Science, Technology, and Humanities. Wiley, New York (1979)
8. Hernandez-Alvarez, M., Gomez, J.M.: Survey about citation context analysis: Tasks, techniques, and resources. Natural Language Engineering **22**(3), 327–349 (2016)
9. Hernndez-Alvarez, M., Gomez Soriano, J.M., Martinez-Barco, P.: Citation function, polarity and influence classification. Natural Language Engineering **23**(4), 561–588 (2017)
10. Kim, I., Thoma, G.R.: Machine learning with selective word statistics for automated classification of citation subjectivity in online biomedical articles. In: Proc. of ICAI'17. pp. 201–207 (2017)
11. Liu, Y., Yan, R., Yan, H.: Guess what you will cite: Personalized citation recommendation based on users' preference. In: Banchs, R.E.e.a. (ed.) IR Technology, LNCS, Volume 8281. pp. 428–439. Springer, Berlin, Heidelberg (2013)
12. Lykke, M., Larsen, B., Lund, H., Ingwersen, P.: Developing a test collection for the evaluation of integrated search. In: Gurrin, C.e.a. (ed.) Advances in Information Retrieval. pp. 627–630. Springer, Berlin, Heidelberg (2010)
13. Mei, Q., Zhai, C.: Generating impact-based summaries for scientific literature. In: In Proc. of ACL'08 HLT. pp. 816–824. ACL (2008)

14. O'Connor, J.: Citing statements: Computer recognition and use to improve retrieval. Information Processing & Management **18**(3), 125 – 131 (1982)
15. Qazvinian, V., Radev, D.R.: Scientific paper summarization using citation summary networks. In: Proc. of the 22Nd ICCL - Volume 1. pp. 689–696. ACL, Stroudsburg, PA, USA (2008)
16. Ritchie, A.: Citation context analysis for information retrieval. University of Cambridge Computer Laboratory Technical Report 744 (2009)
17. Roberts, K., Demner-Fushman, D., Voorhees, E.M., Hersh, W.R.: Overview of the trec 2016 clinical decision support track. In: Proc. of TREC 25. pp. 1–14 (2017)
18. Small, H.: Citation context analysis. In: Dervin, P., Voigt, M.J. (eds.) Progress in Communication Sciences, pp. 287–310. Ablex, Norwood, N.J.
19. Small, H.: Interpreting maps of science using citation context sentiments: a preliminary investigation. Scientometrics **87**(2), 373–388 (2011)
20. Small, H., Tseng, H., Patek, M.: Discovering discoveries: Identifying biomedical discoveries using citation contexts. Journal of Informetrics **11**(1), 46–62 (2017)
21. Strohman, T., Metzler, D., Turtle, H., Croft, W.B.: Indri: a language-model based search engine for complex queries. Tech. rep., In Proc. of the International Conference on Intelligent Analysis (2005)
22. Teufel, S., Siddharthan, A., Tidhar, D.: Automatic classification of citation function. In: Proc. of EMNLP'06. pp. 103–110. ACL, Stroudsburg, PA, USA (2006)
23. White, H.D.: Citation analysis and discourse analysis revisited. Applied Linguistics **25**(1), 89–116 (2004)
24. Zorita, C.H., Moreno-Sandoval, A.: Sentence length and np complexity of general and medical written academic and media texts. an analysis using a trained syntactic parser. In: Ortiz, A.M., Perez-Hernandez, C. (eds.) In Proc. of CILC2016. 8th International Conference on Corpus Linguistics. EPiC Series in Language and Linguistics, vol. 1, pp. 181–190 (2016)