Citation Resolution: A method for evaluating context-based citation recommendation systems

Daniel Duma
University of Edinburgh
D.C.Duma@sms.ed.ac.uk

Ewan Klein
University of Edinburgh
ewan@staffmail.ed.ac.uk

Abstract

Wouldn't it be helpful if your text editor automatically suggested papers that are relevant to your research? Wouldn't it be even better if those suggestions were contextually relevant? In this paper we name a system that would accomplish this a *context-based citation recommendation (*CBCR*) system.* We specifically present Citation Resolution, a method for the evaluation of CBCR systems which exclusively uses readily-available scientific articles. Exploiting the human judgements that are already implicit in available resources, we avoid purpose-specific annotation. We apply this evaluation to three sets of methods for representing a document, based on a) the contents of the document, b) the surrounding contexts of citations to the document found in other documents, and c) a mixture of the two.

1 Introduction

Imagine that you were working on a draft paper which contained a sentence like the following:[1]

> A variety of coherence theories have been developed over the years ... and their principles have found application in many symbolic text generation systems (e.g. [CITATION HERE])

Wouldn't it be helpful if your editor automatically suggested some references that you could cite here? This is what a citation recommendation system ought to do. If the system is able to take into account the *context* in which the citation occurs — for example, that papers relevant to our example above are not only about text generation systems, but specifically mention applying coherence theories — then this would be much more informative. So we define a *context-based* citation recommendation (CBCR) system as one that assists the author of a draft document by suggesting other documents with content that is relevant to a particular context in the draft.

Our longer term research goal is to provide suggestions that satisfy the requirements of specific expository or rhetorical tasks, e.g. provide support for a particular argument, acknowledge previous work that uses the same methodology, or exemplify work that would benefit from the outcomes of the author's work. However, our current paper has more modest aims: we present initial results using existing IR-based approaches and we introduce an evaluation method and metric. CBCR systems are not yet widely available, but a number of experiments have been carried out that may pave the way for their popularisation, e.g. He et al. (2010), Schäfer and Kasterka (2010) and He et al. (2012). It is within this early wave of experiments that our work is framed.

A main problem we face is that evaluating the performance of these systems ultimately requires human judgement. This can be captured as a set of relevance judgements for candidate citations over a corpus of documents, which is an arduous effort that requires considerable manual input and very careful preparation. In designing a context-based citation recommendation system, we would ideally like to minimise these costs.

Fortunately there is already an abundance of data that meets our requirements: every scientific paper contains human "judgements" in the form of citations to other papers which are contextually appropriate: that is, relevant to specific passages of the document and aligned with its argumentative structure. Citation Resolution is a method for evaluating CBCR systems that is exclusively based on this source of human judgements.

[1] Adapted from the introduction to Barzilay and Lapata (2008)

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 358–363,
Baltimore, Maryland, USA, June 23-25 2014. ©2014 Association for Computational Linguistics

Let's define some terminology. In the following passage, the strings 'Scott and de Souza, 1990' and 'Kibble and Power, 2004' are both *citation tokens*:

> A variety of coherence theories have been developed over the years ... and their principles have found application in many symbolic text generation systems (e.g. Scott and de Souza, 1990; Kibble and Power, 2004)

Note that a citation token can use any standard format. Furthermore

- a *citation context* is the context in which a citation token occurs, with no limit as to representation of this context, length or processing involved;
- a *collection-internal reference* is a reference in the bibliography of the source document that matches a document in a given corpus;
- a *resolvable citation* is an in-text citation token which resolves to a collection-internal reference.

2 Related work

While the existing work in this specific area is far from extensive, previous experiments in evaluating context-based citation recommendation systems have used one of three approaches. First, evaluation can be carried out through user studies, which is costly because it cannot be reused (e.g. Chandrasekaran et al. (2008)).

Second, a set of relevance judgements can be created for repeated testing. Ritchie (2009) details the building of a large set of relevance judgements in order to evaluate an experimental document retrieval system. The judgements were mainly provided by the authors of papers submitted to a locally organised conference, for over 140 queries, each of them being the main research question of one paper. This is a standard approach in IR, known as building a *test collection* (Sanderson, 2010), which the author herself notes was an arduous and time-consuming task.

Third, as we outlined above, existing citations between papers can be exploited as a source of human judgements. The most relevant previous work on this is He et al. (2010), who built an experimental CBCR system using the whole index of CiteSeerX as a test collection (over 450,000 documents). They avoided direct human evaluation and instead used three relevance metrics:

- *Recall*, the presence of the original reference in the list of suggestions generated by the system;
- *Co-cited probability*, a ratio between, on the one hand, the number of papers citing both the original reference and a recommended one, and on the other hand, the number of papers citing either of them; and
- *Normalized Discounted Cumulative Gain*, a measure based on the rank of the original reference in the list of suggested references, its score decreasing logarithmically.

However, these metrics fail to adequately recognise that the particular reference used by an author e.g. in support of an argument or as exemplification of an approach, may not be the most appropriate that could be found in the whole collection. This does not just amount to a difference of opinion between different authors; it is possible that within a large enough collection there exists a paper which the original author herself would consider to be more appropriate by any criteria (persuasive power, discoverability or the publication, etc.) than the one actually cited in the paper. Also, given that recommending the original citation used by the author in first position is our key criterion, a metric with smooth discounting like NDCG is too lenient for our purposes.

We have then chosen *top-1 accuracy* as our metric, where every time the original citation is first on the list of suggestions, it receives a score of 1, and 0 otherwise, and these scores are averaged over all resolved citations in the document collection. This metric is intuitive in measuring the efficiency of the system at this task, as it is immediately interpretable as a percentage of success.

While previous experiments in CBCR, like the ones we have just presented, have treated the task as an Information Retrieval problem, our ultimate purpose is different and travels beyond IR into Question Answering. We want to ultimately be able to assess the reason a document was cited in the context of the argumentation structure of the document, following previous work on the automatic classification of citation function by Teufel et al. (2006), Liakata et al. (2012) and Schäfer and Kasterka (2010). We expect this will allow us to identify claims made in a draft paper and match them with related claims made in other papers for support or contrast, and so offer answers in the form of relevant passages extracted from the suggested documents.

It is frequently observed that the reasons for citing a paper go beyond its contribution to the field and its relevance to the research being reported (Hyland, 2009). There is a large body of research on the motivations behind citing documents (MacRoberts and MacRoberts, 1996), and it is likely that this will come to play a part in our research in the future.

In this paper, however, we present our initial results which compare three different sets of IR-based approaches to generating the document representation for a CBCR system. One is based on the contents of the document itself, one is based on the existing contexts of citations of this paper in other documents, and the third is a mixture of the two.

3 The task: Citation Resolution

In this section we present the evaluation method in more abstract terms; for the implementation used in this paper, please see Sections 4 and 5. The core criterion of this task is to use only the human judgements that we have clearest evidence for. Let d be a document and R the collection of all documents that are referenced in d. We believe it is reasonable to assume that the author of document d knows enough about the contents of each document R_i to choose the most appropriate citation from the collection R for every citation context in the document.

This captures a very strong relevance judgement about the relation between a particular citation context in the document and a particular cited reference document. We use these judgements for evaluation: our task is to match every citation context in the document (i.e. the surrounding context of a citation token) with the right reference from the list of references cited by that paper.

This task differs somewhat from standard Information Retrieval, in that we are not trying to retrieve a document from a larger collection outside the source document, but trying to resolve the correct reference for a given citation context from an existing list of documents, that is, from the bibliography that has been manually curated by the authors. Our document collection used for retrieval is further composed of only the references of that document that we can access.

The algorithm for the task is presented in Figure 1. For any given *test document* (2), we first extract all the citation tokens found in the text that correspond to a collection-internal reference (a). We then create a *document representation* of the referenced document (currently a Vector Space Model, but liable to change). This representation can be based on any information found in the document collection, excluding the document d itself: e.g. the text of the referenced document and the text of documents that cite it.

For each citation token we then extract its context (b.i), which becomes the *query* in IR terms. One way of doing this that we present here is to select a list of word tokens around the citation. We then attempt to *resolve* the citation by computing a score for the match between each reference representation and the citation context (b.ii). We rank all collection-internal references by this score in decreasing order, aiming for the original reference to be in the first position (b.iii).

In the case where multiple citations share the same context, that is, they are made in direct succession (e.g. *"...compared with previous approaches (Author (2005), Author and Author (2007))"*), the first n elements of the list of suggested documents all count as the first element. That is, if any of the references in a multiple citation of n elements appears in the first n positions of the list of suggestions, it counts as a successful resolution and receives a score of 1. The final score is averaged over all citation contexts processed.

The set of experiments we present here apply this evaluation to test a number of IR techniques which we detail in the next section.

```
1. Given document collection $D$
2. For every test document $d$
    (a) For every reference $r$ in its bibliography $R$
        i. If $r$ is in document collection $D$
        ii. Add all inline citations $C_r$ in $d$ to list $C$
    (b) For each citation $c$ in $C$
        i. Extract context $ctx_c$ of $c$
        ii. Choose which document $r$ in $R$ best matches $ctx_c$
        iii. Measure accuracy
```

Figure 1: Algorithm for citation resolution.

4 Experiments

Our test corpus consists of approx. 9000 papers from the ACL Anthology [2] converted from PDF to

[2] http://http://aclweb.org/anthology/

XML format. This corpus, the rationale behind its selection and the process used to convert the files is described in depth in Ritchie et al. (2006). This is an ideal corpus for these tests for a large number of reasons, but these are key for us: all the papers are freely available, the ratio of collection-internal references for each paper is high (the authors measure it at 0.33) and it is a familiar domain for us.

For our tests, we selected the documents of this corpus with at least 8 collection-internal references. This yielded a total of 278 test documents and a total of 5446 resolvable citations.

We substitute all citations in the text with citation token placeholders and extract the citation context for each using a simple *window* of up to w words left and w words right around the placeholder. This produces a list of word tokens that is equivalent to a *query* in IR.

This is a frequently employed technique (He et al., 2010), although it is often observed that this may be too simplistic a method (Ritchie, 2009). Other methods have been tried, e.g. full sentence extraction (He et al., 2012) and comparing these methods is something we plan to incorporate in future work.

We then make the document's collection-internal references our test collection D and use a number of methods for generating the document representation. We use the well-known Vector Space Model and a standard implementation of *tf-idf* and *cosine similarity* as implemented by the *scikit-learn* Python framework [3]. At present, we are applying no cut-off and just rank all of the document's collection-internal references for each citation context, aiming to rank the correct one in the first positions in the list.

We tested three different approaches to generating a document's VSM representation: *internal representations*, which are based on the contents of the document, *external representations*, which are built using a document's incoming link citation contexts (following Ritchie (2009) and He et al. (2010)) and mixed representations, which are an attempt to combine the two.

- The internal representations of the documents were generated using three different methods: title plus abstract, full text and *passage*. Passage consists in splitting the document into half-overlapping passages of a fixed length of k words and choosing for each document the passage with the maximum cosine similarity score with the query. We present the results of using 250, 300 and 350 as values for k.
- The external representations (*inlink_context*) are based on extracting the context around citation tokens to the document from other documents in the collection, excluding the set of test papers. This is the same as using the *anchor text* of a hyperlink to improve results in web-based IR (see Davison (2000) for extensive analyis). This context is extracted in the same way as the query: as a window, or list of w tokens surrounding the citation left and right. We present our best results, using symmetrical and asymmetrical windows of $w = [(5,5),(10,10),(10,5),(20,20),(30,30)]$.
- We build the mixed representations by simply concatenating the internal and external bags-of-words that represent the documents, from which we then build the VSM representation. For this, we combine different window sizes for the *inlink_context* with: *full_text*, *title_abstract* and *passage350*.

5 Results and discussion

Table 1 presents a selection of the most relevant results, where the best result and document representation method of each type is highlighted. We present results for the most relevant parameter values, producing the highest scores of all those tested.

From a close look at internal methods, we can see that the *passage* method with $k = 400$ beats both *full_text* and *title_abstract*, suggesting that a more elaborate way of building a document representation should improve results. This is consistent with previous findings: Gay et al. (2005) had already reported that using selected sections plus captions of figures and title and abstract to build the internal document representation improves the results of their indexing task by 7.4% over just using title and abstract. Similarly, Jimeno-Yepes et al. (2013) showed that automatically generated summaries lead to similar recall and better indexing precision than full-text articles for a keyword-based indexing task.

However, it is immediately clear that purely external methods obtain higher scores than internal ones. The best score of 0.413 is obtained by the *inlink_context* method with a window of 10 tokens left, 5 right, combined with the similarly-sized ex-

[3] http://scikit-learn.org

Method	window5_5	window10_10	window10_5	window20_20	window30_30
Internal methods					
full_text	0.318	0.340	0.337	0.369	0.370
title_abstract	0.296	0.312	0.312	0.322	0.311
passage250	0.343	0.367	0.359	0.388	0.382
passage350	0.346	0.371	0.364	0.388	0.381
passage400	0.348	0.371	0.362	**0.391**	0.380
External methods					
inlink_context10	0.391	0.406	0.405	0.395	0.387
inlink_context20	0.386	0.406	**0.413**	0.412	0.402
inlink_context30	0.380	0.403	0.400	0.411	0.404
Mixed methods					
inlink_context_20_full_text	0.367	0.407	0.399	0.431	0.425
inlink_context_20_title_abstract	0.419	0.447	0.441	0.453	0.437
inlink_context_20_passage250	0.420	0.458	0.451	**0.469**	0.451
inlink_context_10_passage350	0.435	0.465	0.459	0.464	0.450
inlink_context_20_passage350	0.426	0.464	0.456	**0.469**	0.456

Table 1: Accuracy for each document representation method (rows) and context window size (columns).

traction method for the query (*window10_10*). We find it remarkable that *inlink_context* is superior to internal methods, beating the best (*passage400*) by 0.02 absolute accuracy points. Whether this is because the descriptions of these papers in the contexts of incoming link citations capture the essence or key relevance of the paper, or whether this effect is due to authors reusing their work or to these descriptions originating in a seed paper and being then propagated through the literature, remain interesting research questions that we intend to tackle in future work.

The key finding from our experiments is however that a mixture of internal and external methods beats both individually. The highest score is 0.469, achieved by a combination of *inlink_context_20* and the *passage* method, for a window of $w = 20$, with a tie between using 250 and 350 as values for k (passage size). The small difference in score between parameter values is perhaps not as relevant as the finding that, taken together, mixed methods consistently beat both external and internal methods.

These results also show that the task is far from solved, with the highest accuracy achieved being just under 47%. There is clear room for improvement, which we believe could firstly come from a more targeted extraction of text, both for generating the document representations and for extracting the citation contexts.

Our ultimate goal is matching claims and comparing methods, which would likely benefit from an analysis of the full contents of the document and not just previous citations of it, so in future work we also intend to use the context from the successful external results as training data for a summarisation stage.

6 Conclusion and future work

In this paper we have presented Citation Resolution: an evaluation method for context-based citation recommendation (CBCR) systems. Our method exploits the implicit human relevance judgements found in existing scientific articles and so does not require purpose-specific human annotation.

We have employed Citation Resolution to test three approaches to building a document representation for a CBCR system: internal (based on the contents of the document), external (based on the surrounding contexts to citations to that document) and mixed (a mixture of the two). Our evaluation shows that: 1) using chunks of a document (passages) as its representation yields better results that using its full text, 2) external methods obtain higher scores than internal ones, and 3) mixed methods yield better results than either in isolation.

We intend to investigate more sophisticated ways of document representation and of extracting a citation's context. Our ultimate goal is not just to suggest to the author documents that are "relevant" to a specific chunk of the paper (sentence, paragraph, etc.), but to do so with attention to rhetorical structure and thus to citation function. We also aim to apply our evaluation to other document collections in different scientific domains in order to test to what degree these results can be generalized.

References

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Kannan Chandrasekaran, Susan Gauch, Praveen Lakkaraju, and Hiep Phuc Luong. 2008. Concept-based document recommendations for citeseer authors. In *Adaptive Hypermedia and Adaptive Web-Based Systems*, pages 83–92. Springer.

Brian D Davison. 2000. Topical locality in the web. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 272–279. ACM.

Clifford W Gay, Mehmet Kayaalp, and Alan R Aronson. 2005. Semi-automatic indexing of full text biomedical articles. In *AMIA Annual Symposium Proceedings*, volume 2005, page 271. American Medical Informatics Association.

Qi He, Jian Pei, Daniel Kifer, Prasenjit Mitra, and Lee Giles. 2010. Context-aware citation recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 421–430. ACM.

Jing He, Jian-Yun Nie, Yang Lu, and Wayne Xin Zhao. 2012. Position-aligned translation model for citation recommendation. In *String Processing and Information Retrieval*, pages 251–263. Springer.

Ken Hyland. 2009. *Academic discourse: English in a global context*. Bloomsbury Publishing.

Antonio J Jimeno-Yepes, Laura Plaza, James G Mork, Alan R Aronson, and Alberto Díaz. 2013. Mesh indexing based on automatically generated summaries. *BMC bioinformatics*, 14(1):208.

Maria Liakata, Shyamasree Saha, Simon Dobnik, Colin Batchelor, and Dietrich Rebholz-Schuhmann. 2012. Automatic recognition of conceptualization zones in scientific articles and two life science applications. *Bioinformatics*, 28(7):991–1000.

Michael H MacRoberts and Barbara R MacRoberts. 1996. Problems of citation analysis. *Scientometrics*, 36(3):435–444.

Anna Ritchie, Simone Teufel, and Stephen Robertson. 2006. Creating a test collection for citation-based ir experiments. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 391–398. Association for Computational Linguistics.

Anna Ritchie. 2009. Citation context analysis for information retrieval. Technical report, University of Cambridge Computer Laboratory.

Mark Sanderson. 2010. *Test collection based evaluation of information retrieval systems*. Now Publishers Inc.

Ulrich Schäfer and Uwe Kasterka. 2010. Scientific authoring support: A tool to navigate in typed citation graphs. In *Proceedings of the NAACL HLT 2010 workshop on computational linguistics and writing: Writing processes and authoring aids*, pages 7–14. Association for Computational Linguistics.

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006. Automatic classification of citation function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 103–110. Association for Computational Linguistics.