High accuracy citation extraction and named entity recognition for a heterogeneous corpus of academic papers

Brett Powley and Robert Dale
Centre for Language Technology
Macquarie University
Sydney, NSW 2109, Australia
{bpowley,rdale}@ics.mq.edu.au

Abstract

Citation indices are increasingly being used not only as navigational tools for researchers, but also as the basis for measurement of academic performance and research impact. This means that the reliability of tools used to extract citations and construct such indices is becoming more critical; however, existing approaches to citation extraction still fall short of the high accuracy required if critical assessments are to be based on them. In this paper, we present techniques for high accuracy extraction of citations from academic papers, designed for applicability across a broad range of disciplines and document styles. We integrate citation extraction, reference parsing, and author named entity recognition to significantly improve performance in citation extraction, and demonstrate this performance on a cross-disciplinary heterogeneous corpus. Applying our algorithm to previously unseen documents, we demonstrate high F-measure performance of 0.98 for author named entity recognition and 0.97 for citation extraction.

1 Introduction

A defining feature of academic literature is the use of citations. Researchers use citations to acknowledge and refer to other works which are related in some way to their own work or their discussion of it. Analysing the relationship between documents implied by citations has long been of interest to researchers; the potential for automation of this analysis was first identified by Garfield in 1955 (Garfield, 1955). While citation indices of the type proposed by Garfield and implemented in the ISI Web of Science (founded by Garfield himself), and more recently in prototype web-based services such as CiteSeer and Google Scholar, provide useful navigation tools to researchers, the increasing use of citation counts as a measure of research impact for the purposes of assessing academic performance means that accurate extraction of citations is becoming more critical. To have confidence in such measures of performance, we need to know that the algorithm used for extraction of the citations on which such counts are based performs with high accuracy, across the full range of academic literature; existing reported work (Bergmark, 2000; Bergmark et al., 2001; Besagni et al., 2003; Giuffrida et al., 2000; Seymore et al., 1999; Takasu, 2003) falls short of this high accuracy. Our focus in this paper is on development of such a high-accuracy algorithm, and in particular on assessing its performance on a wide variety of document formats.

We use the terms *citation* and *reference* as follows: a *reference* appears in a list of works at the end of a document, and provides full bibliographic information about a cited work; a *citation* is a mention of a work in the body of the text, and includes enough information (typically, an author–year pair or an alphanumeric key) to uniquely identify the work in the list of references.

Powley and Dale (2007) introduced terminology to describe the variety of citation styles encountered in academic literature; Table 1 summarises the major styles. For the present work, we are interested in *textual citations* only: they are more relevant to our related work on analysing the function of cita-

978-1-4244-1610-3/07/$25.00©2007IEEE

Textual - Syntactic **Levin (1993)** provides a classification of over 3000 verbs according to their participation in alternations ...
Textual - Parenthetical Two current approaches to English verb classifications are WordNet (Miller et al., 1990) and Levin classes **(Levin, 1993)**.
Prosaic **Levin** groups verbs based on an analysis of their syntactic properties ...
Pronominal **Her** approach reflects the assumption that the syntactic behavior of a verb is determined in large part by its meaning.
Numbered There are patterns of diathesis behaviour among verb groups **[1]**.

Table 1: Citation styles

tions; and in many ways they present the more difficult case, so the work involved in extracting textual citations is a superset of that required to work on indexed citations.

In earlier work (Powley and Dale, 2007), we used an integrated evidence-based algorithm for extraction of citations from the the Association for Computational Linguistics Anthology[1], a digital archive of approximately 10,000 conference and journal papers in computational linguistics. Although the corpus includes documents from a range of conferences, workshops, and journals, it is still evident that the variety of document styles in such a corpus is limited. In this work, we aim to extend the algorithm's effectiveness to a broad range of document styles.

2 Related work

There have been several approaches to the problem of extracting citations from academic literature, frequently on corpora consisting of documents from a common source. Bergmark et al. (2001) report on heuristics for extracting citations from ACM papers, reporting precision of 0.53, based on randomly selected papers. Bergmark (2000) reports in more detail on extracting information from digital library papers, including citations in a variety of formats, reporting 86.1% 'average accuracy' for elements extracted from each document; she does not report performance for citation extraction separately. Besagni et al. (2003) use part-of-speech tagging of words in references on a corpus of pharmacology journal papers, and report 90.2% accuracy in extracting author names. Takasu (2003) employs hidden Markov models and support vector machines for reference segmentation in a corpus of (English-language) Japanese computer science journals, reporting high accuracy results, but also pointing out that their test corpus had extremely consistent formatting. Giuffrida et al. (2000) use a knowledge-based system to extract metadata from computer science journal papers, reporting 87% accuracy in extracting author names. Seymore et al. (1999) use hidden Markov models on a similar corpus, reporting 93.2% accuracy for author name extraction.

3 The heterogeneous corpus

In earlier work (Powley and Dale, 2007), we evaluated our citation extraction algorithm on documents drawn from the various styles of paper (conferences, workshops, journals) in the ACL Anthology. For the present work, we wanted to generalise our approach and evaluate its performance on a corpus drawn from outside this relatively narrow field. As a source for documents, we used IngentaConnect[2], an online repository representing around 30,000 academic publications across a full range of disciplines. To extract a random selection of articles, we searched using a term very commonly found in abstracts, the word *show* (the Ingenta search engine disallows searches on common stopwords such as *the*). We generated a list of all articles containing this word in a 12 month period from January 2006 to January 2007, restricted our results to those for which our institution had full-text access, and further restricted the collection to those containing textual citations, since these are the focus of our work. We then chose at random a single document from each journal, yielding a corpus of 216 documents. The distribution of the documents across disciplines is shown in Table 2.

[1]The ACL Anthology, available at http://acl.ldc.upenn.edu/.

[2]http://www.ingenta.com

Astronomy	1
Bioscience	45
Economics	34
Education	25
Geoscience	24
History & Politics	10
Informatics	8
Linguistics	8
Mathematics & Statistics	12
Medicine & Health	15
Philosophy	2
Psychology	25
Social Sciences	21

Table 2: The heterogeneous corpus

From this corpus, we chose at random 50 documents as a training corpus for refining our citation extraction heuristics, and withheld a separate 50 documents as a test corpus for the experiments in this paper.

4 Document preprocessing

The source documents in our corpus are in PDF format. To produce a text corpus for processing, we used an open-source tool[3] to extract text from the PDF sources. Our current work focusses on techniques which will work on an unformatted text stream; the only intact formatting cues from the source document are line breaks, with font changes, blank lines, and all other formatting absent. We have chosen relatively recent documents for our corpus since they are more likely to be 'born digital' rather than scanned; this allows us to isolate our algorithm's performance from OCR errors in older scanned documents.

Individual documents are segmented into header, body, references and appendix sections using textual cues. Lines containing the copyright symbol © or *(c)* are discarded, as this simple heuristic identifies a large number of unwanted page headers and footers. The body section of the document is then dehyphenated and segmented into sentences. The data to be processed for each document then comprises a list of sentences from the body of the document, and a segment of text representing the references section.

[3] PDFBox, available at http://www.pdfbox.org/

```
<citation-instance> ::= <author-list> <words>* <year-list>
<words> ::= non-author words
<author-list> ::= { <author-surname> <author-separator>* }+ [et al]['s]
<author-separator> ::= , | ; | and | &
<year-list> ::= [ ( ] { <year> <year-separator>* }+ [ ) ]
<year-separator> ::= , | ;
<year> ::= { 1900 | 1901 | 1902 | ... | current year } [ a | b | c | ... ]
```

Figure 1: Simplified grammar for citations

5 Citation extraction

5.1 Algorithm

The citation extraction algorithm works at the sentence level to isolate and tag citations. We begin with the observation that textual citations are anchored around years; we previously showed that we could identify candidate sentences containing citations with a recall of better than 0.99 simply by using the presence of a year as a cue (Powley and Dale, 2007).

Our first step is therefore to search each sentence for a candidate year token (a 'year' for this purpose being a 4-digit number between 1900 and the current year, potentially with a single character appended to it). If we find such a token, our task is then to determine whether it forms part of a citation, and if it does, to extract the author names that accompany it. A simplified version of the grammar for a citation on which our algorithm is based is shown in Figure 1.

In general, we may say that a textual citation comprises one or more authors followed by one or more years; in practice, the variety of constructions which a writer might use to format a citation is somewhat more complicated.

Writers often use a list of years as shorthand for citing multiple papers by the same author: consider *Smith (1999; 2000)*, which represents citations of two separate works. Given the candidate year, we therefore first search backwards and forwards to isolate a list of years. Our task is then to find the list of authors. While this often immediately precedes the list of years, this is not always the case; consider, for example, *Knuth's prime number algorithm (1982)*. We therefore search backwards from the year list, skipping words until we find an author name; currently, we choose to stop searching after 10 words,

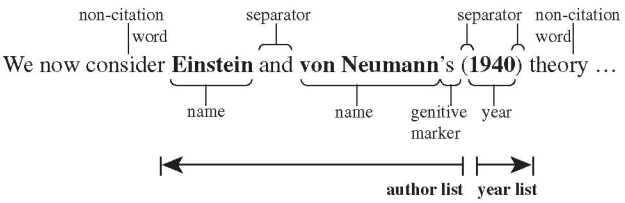

Figure 2: Extracting citation information

as we have found that this choice gives good performance. Having found a single author name, we continue searching backwards for additional names, skipping over punctuation and separators, and stopping when we encounter a non-surname word; an illustration of this process is shown in Figure 2. If no author names are found, we conclude that the candidate year was a number unrelated to a citation. We also treat a small number of temporal prepositions which commonly appear before years as stopwords, concluding that the candidate year is not a citation if preceded by one of these (*in*, *since*, *during*, *until*, and *before*). Otherwise, having found a list of authors, we normalise the citation instance into a list of citations each consisting of a list of authors and a single year. We also record whether the citation contains an *et al* string (indicating that the list of authors is not comprehensive) or whether it ends with a genitive (e.g. *Powley's (2006) citation extraction algorithm*). The key problem in citation extraction is accurate identification of author surnames, the algorithm for which is described in the following section.

5.2 Evaluation and Error Analysis

To evaulate our algorithm, we ran the citation extractor on the heterogeneous test corpus, producing a list of candidate citing sentences (those containing years), and a list of citations found in each sentence. This output was then hand annotated to tag (a) words incorrectly identified as citations; (b) correctly identified but incorrectly segmented citations; and (c) missed citations. Where a citation was missed, we also tagged whether the citation had any errors (misspelled names, incorrect years, or a missing corresponding reference). The results are shown in Table 3, along with the results of our earlier experiments (Powley and Dale, 2007) for comparison.

The high precision of the algorithm is largely due to the high performance of the named entity recognition algorithm on which it relies, described in Section 6. While recall performance is still good, it is lower than we achieved with the ACL Anthology corpus.

	Heterogeneous corpus	Anthology corpus
Number of documents	50	60
Citation instances	4011	2406
Precision	0.9904	0.9992
Recall	0.9416	0.9612
F-measure	0.9712	0.9798

Table 3: Citation extraction results

Examining the 'missed' citations, we find that 30% of the citations missed by our algorithm on this corpus are from just three documents. In two of these, the document segmentation task failed to recognise the references section, in one because the author had used a novel and unusual name for it; and in the other because the references section had no header at all. Poor performance on the third document appeared to be the result of unusually bad editing: 22% of the citations in this document had no corresponding reference. If we exclude the first two atypical documents from our results, the resulting recall of 0.9527 is close to the performance on the Anthology corpus.

The issue of writer errors warrants closer study; isolating those citations missed due to misspelled names, incorrect years, or missing references reveals that 38% of the missed citations were due to our algorithm not handling such errors. Across the test corpus, 26% of articles contained at least one error, 10% had five or more errors, and two particularly egregious examples contained 20 or more errors. This suggests that a significant target for improvement in our algorithm's performance ought to be developing strategies for handling degenerate data, and in particular for isolating and resolving citations when no matching reference is found using strict matching.

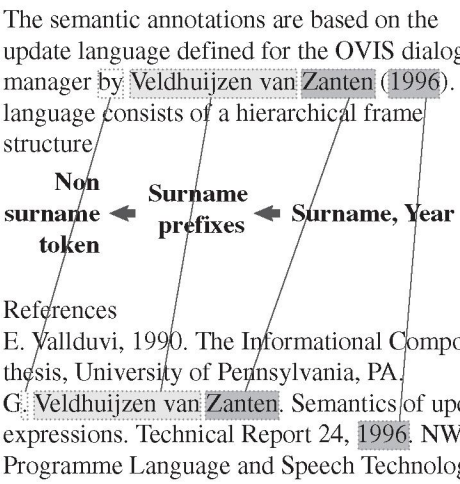

Figure 3: Named entity recognition

	Heterogeneous		Anthology	
Author named entity recognition				
	A	B	A	B
Precision	0.99	0.75	1.0	0.92
Recall	0.98	0.97	0.97	1.0
F-measure	0.98	0.85	0.98	0.96
Author prefix identification				
	A	B	A	B
Precision	1.0	0.09	0.92	0.36
Recall	0.96	0.51	1.0	0.26
F-measure	0.98	0.15	0.96	0.30

Table 4: Named entity recognition for Alignment (A) and Baseline (B) algorithms

6 Alignment-based named entity recognition

The citation extraction algorithm relies on the ability to identify author surnames. In particular, we require the ability to find the author name list preceding a candidate year by distinguishing author names from other words.

Our named entity recognition algorithm is based on the observation that any author name in the body of the document ought also to appear in the references section. A candidate surname in a citation is a capitalised token preceding a year (or another surname); the simplest approach is to search the references section for the same token, and if it appears, assume that the candidate token is a surname. However, surnames are not the only capitalised words that appear in the references section, so treating the entire references section simply as an author name gazetteer in this fashion generates false positives. To be more certain that the token that we have found in the references section is an author name, we use an alignment algorithm, searching for both the candidate surname and the year from the reference appearing within a 5-line window of each other.

An additional problem that we want our named entity recogniser to be able to handle is that of compound surnames, or surnames which consist of more than a single capitalised word; we have found that approximately 2% of author names in a moderately-sized corpus are of this type (Powley and Dale, 2007). Commonly, these comprise a surname and a number of prefixes (often prepositions) from a relatively fixed set: for example, *von Neumann*, *van den Bosch*, *Uit den Boogaart*, *Della Pietra*, and *Al Shalabi*. Our initial approach to this problem was therefore to build a baseline algorithm based on an authoritative list of surname prefixes, and tag items from this list preceding a capitalised surname as part of the surname. However, experiments on both the ACL Anthology and the heterogeneous corpus showed poor performance using this simple baseline; results are shown in Table 4. The main reason for this is that compound surnames which consist of elements from a non-closed set are not uncommon: for example, *Gaustad van Zaanen*, *Villemonte de la Clergerie*, *Tjong Kim Sang*, and *Schulte im Walde*.

Our strategy for detecting the bounds of compound surnames is then to make no assumptions about the set of words which can comprise a surname and its prefixes. Rather, we use evidence from the two (or more) distinct instances of a surname which we have: one in the body of the document as part of the citation, and one in the references section as part of a reference. An example of compound named entity recognition is shown in Figure 3. We start at the capitalised surname word in the body text, and its counterpart in the references text. Moving backwards in the body and references, we compare words, continuing until a non-matching word is found. Matching words are then tagged as part of the surname.

The results of running the alignment-based algorithm on the heterogeneous corpus are shown in Table 4, with the performance of the baseline algorithm and earlier results on the Anthology corpus shown for comparison. The alignment-based algorithm gives extremely good results: the high precision (0.99) shows that we rarely misidentify a token as an author name, or miss author name prefixes. The high recall indicates that we rarely miss author names; error analysis shows that the main cause of missed names in our test data set was malformed author names or missing references.

7 Discussion and future work

The citation extraction algorithm introduced in Powley and Dale (2007) performs well on our heterogeneous corpus, with results comparable to those we reported earlier on the ACL Anthology corpus; this validates our evidence-based approach to citation extraction and named entity recognition as being more broadly applicable.

While this work (and our related work on citation analysis) has focussed on textual citations, we plan to extend this approach to indexed citations in future work; construction of a broad-coverage heterogeneous corpus similar to that constructed for this work will be a key part of evaluating that work.

The high incidence of referencing errors even in published journal articles was an important finding in this work. This was somewhat surprising, since one might otherwise assume that the majority of academic documents would be authored using bibliographic tools such as BibTEX or Endnote, limiting the potential for such errors; and further that the review and editing process for journal papers should pick up the majority of such errors. Given that this appears not to be the case, an important direction for citation extraction work will be the detection and handling of citation and reference errors. Tools which can handle such errors will be more useful not only for higher performance citation extraction, but also have a more immediate practical application: automatically detecting, reporting on, and correcting errors, perhaps as part of automatically checking papers submitted to a conference or journal.

References

Donna Bergmark. Automatic extraction of reference linking information from online documents. Technical Report CSTR2000-1821, Cornell Digital Library Research Group, 2000.

Donna Bergmark, Paradee Phempoonpanich, and Shumin Zhao. Scraping the ACM digital library. *SIGIR Forum*, 35(2), 2001. URL `http://www.acm.org/sigir/forum/F2001-TOC.html`.

D. Besagni, A. Belaid, and N. Benet. A segmentation method for bibliographic references by contextual tagging of fields. In *Document Analysis and Recognition, 2003. Proceedings. Seventh International Conference on, Vol., Iss., 3-6 Aug. 2003*, pages 84– 388 vol.1, 2003.

Eugene Garfield. Citation indexes for science: A new dimension in documentation through association of ideas. *Science*, 122(3159):108–111, July 1955.

Giovanni Giuffrida, Eddie C. Shek, and Jihoon Yang. Knowledge-based metadata extraction from postscript files. In *DL '00: Proceedings of the fifth ACM conference on Digital libraries*, pages 77–84. ACM Press, 2000. ISBN 1-58113-231-X. doi: http://doi.acm.org/10.1145/336597.336639.

Brett Powley and Robert Dale. Evidence-based information extraction for high-accuracy citation extraction and author name recognition. In *Proceedings of the 8th RIAO International Conference on Large-Scale Semantic Access to Content*, Pittsburgh, PA, 2007.

Kristie Seymore, Andrew McCallum, and Roni Rosenfeld. Learning hidden markov model structure for information extraction. In *AAAI 99 Workshop on Machine Learning for Information Extraction*, 1999.

Atsuhiro Takasu. Bibliographic attribute extraction from erroneous references based on a statistical model. In *Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries*, pages 49–60, 2003. URL `http://portal.acm.org/citation.cfm?id=827147`.