The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics

Steven Bird[1], Robert Dale[2], Bonnie J. Dorr[3], Bryan Gibson[4], Mark T. Joseph[4], Min-Yen Kan[5†], Dongwon Lee[6], Brett Powley[2], Dragomir R. Radev[4], Yee Fan Tan[5]

[1]University of Melbourne, [2]Macquarie University, [3]University of Maryland,
[4]University of Michigan, [5]National University of Singapore, [6]The Pennsylvania State University

sb@csse.unimelb.edu.au, {rdale,bpowley}@ics.mq.edu.au,
bonnie@umiacs.umd.edu, {gibsonb,mtjoseph,radev}@umich.edu,
{kanmy,tanyeefa}@comp.nus.edu.sg, dongwon@psu.edu

Abstract

The ACL Anthology is a digital archive of conference and journal papers in natural language processing and computational linguistics. Its primary purpose is to serve as a reference repository of research results, but we believe that it can also be an object of study and a platform for research in its own right. We describe an enriched and standardized reference corpus derived from the ACL Anthology that can be used for research in scholarly document processing. This corpus, which we call the ACL Anthology Reference Corpus (ACL ARC), brings together the recent activities of a number of research groups around the world. Our goal is to make the corpus widely available, and to encourage other researchers to use it as a standard testbed for experiments in both bibliographic and bibliometric research.

1. Introduction

The advent of scholarly digital libraries has tremendously facilitated access to published research. In many fields, scholars now often use such digital libraries as their entry point into the research literature. Modern digital libraries rely on a number of semi-automated tasks, such as document collection and reference metadata extraction and cleaning, and provide infrastructure for searching and browsing. High performance on these tasks is critical to enabling lightweight, low-cost quality maintenance of a digital collection. As summarized in Table 1, we are witnessing a proliferation of digital libraries from diverse disciplines and domains. However, to the best of our knowledge, there has been little work on building a standard, real-world digital collection testbed that can be used to measure performance on the key infrastructural tasks that have such an impact on the value of these resources. This paper presents an attempt to provide such a testbed.

Sponsored by the Association for Computational Linguistics, the ACL Anthology represents the NLP community's most up-to-date and long-standing freely accessible research repository.[1] At the time of writing, the Anthology contains 14,000 articles, drawn from a range of conferences and workshops as well as past issues of the *Computational Linguistics* journal. It is indexed by a host of other digital libraries and repositories, such as CiteSeer, Google Scholar, OLAC, and the ACM Digital Library. In this paper, we describe the ACL Anthology Reference Corpus (ACL ARC),[2] a collaborative attempt to provide a standardized reference corpus based on the ACL Anthology.

Section 2 provides background information on the ACL Anthology. In Section 3, we give an overview of the ACL ARC as an end-product, and then describe the processing done to the source ACL Anthology data to transform it into the reference corpus. Then we describe future plans for the ACL ARC that include bibliographic processing. In Section 5, we review related work in bibliographic research and discuss how the development of the ACL ARCt relates to recent grassroots initiatives in the community. We conclude our paper with a call to researchers to utilize the ACL ARC as a target corpus in their bibliographic research.

2. Background

The proposal for an ACL Anthology was first put forward to the ACL Executive by Steven Bird at the Association's 2001 conference, in response to a call for ideas to mark the ACL's 40th anniversary. In the following 12 months, over US$50,000 in funding was donated by institutions and individuals to allow digitization of the previous two decades of ACL conference and journal issues. Pages were scanned at 600dpi grayscale for archival storage, and then downsampled to 300dpi black-and-white, and assembled into articles and stored in the 'PDF Image with Hidden Text' format. Author and title metadata was extracted from the OCRed text, and used to build HTML index pages.

By the time of its launch at the 40th anniversary meeting in Philadelphia in 2002, the Anthology contained 3,100 papers, available on the web and indexed by search engines. Later tasks involved locating older materials such as conference proceedings dating back to the 1960s; digitizing microfiche slides from the early years of the journal *Computational Linguistics*; and manually converting the set of 'born-digital' proceedings to the Anthology layout.

[†] Contact author.

[1] Available at http://www.aclweb.org/anthology/.

[2] For ease of reference we use 'ACL ARC' to refer to the corpus project under discussion, and 'ACL Anthology' for the publicly-accessible website containing the ACL publication archives, which currently spans the period from the 1970s to 2007.

Name	Domains	# Articles	# References	Source
ISI SCI	Sciences	0	25m	HH
ISI Science Citation Index http://portal.isiknowledge.com/portal.cgi				
CAS	Chemistry	0	23m	HH
Chemical Abstracts Service http://www.cas.org/				
PubMed	Life Science	0	12m	HH
http://www.ncbi.nlm.nih.gov/sites/entrez				
CiteSeer	Sciences	0.8m	10m	SS
http://citeseer.ist.psu.edu/				
arXiv e-Print	Physics, Math	0.3m	0.3m	HS
http://arxiv.org/				
SPIRES-HEP	High-energy Physics	0.27m	0.5m	HH
SPIRES High Energy Physics Database http://www.slac.stanford.edu/spires/index.shtml/hep/				
DBLP	Computer Science	0	0.93m	H
Digital Bibliography and Library Project http://www.informatik.uni-trier.de/ ley/db/index.html				
CSB	Computer Science	0	2m	SS
Collection of Computer Science Bibliographies http://liinwww.ira.uka.de/bibliography/				
ACM D	Computer Science	N/A	N/A	HS
Association for Computing Machinery Digital Library (Portal) http://portal.acm.org/dl.cfm				
IEEE DL	Engineering	N/A	N/A	HS
Institute of Electrical and Electronic Engineers Digital Library (Xplore) http://www.computer.org/portal/site/csdl/				
SIGMOD Anthology	Computer Science	N/A	N/A	HH
Special Interest Group on the Management of Data Anthology http://www.informatik.uni-trier.de/ ley/db/anthology.html				
Google Scholar	Sciences	N/A	N/A	SS
http://scholar.google.com/				
CNKI	Sciences	0.89m	0.89m	HS (?)
Collection of Full-Text Papers in Important Chinese Conferences (Chinese) http://cajiod.cnki.net/kns50/Navigator.aspx?ID=CPFD				
HKJO	Sciences	N/A	N/A	HH (?)
Hong Kong Journals Online (Chinese and English) http://sunzi1.lib.hku.hk/hkjo/				

Table 1: Demographics of a sample of currently available scholarly digital collections. Sizes are given in millions of PS or PDF articles held by the collection, where 0 indicates that there are no resources available. Source indicates the origin of the data: HH for human-submitted and human-extracted, HS for human-submitted and software-extracted, and SS for software-crawled and software-extracted collections.

Currently, the ACL's conference publication software automatically generates conference proceedings that can be incorporated into the Anthology with a minimum of manual effort. Papers from almost all events sponsored by, or in some way affiliated with, the ACL have now been incorporated into the collection, providing a constantly growing and up-to-date collection of publications in the field.

In recent years, subsets of the Anthology have served as an evaluation corpus for research efforts in bibliographic data processing carried out by researchers in our own community. However, these experiments have employed different subsets of the Anthology at different points in time, making comparisons across experiments difficult. Other research communities, such as those concerned with digital libraries and databases, face the same problem: people often use subsets of reference data from the DBLP or CiteSeer collections, yet the quality of metadata is not satisfactory and there have not been any reference subsets against which research results can be objectively compared. Information Retrieval corpora, such as ones used by TREC and CLEF community evaluations, while standard, largely focus on newswire and are unsuited for bibliographic research. To facilitate future work, a standardized reference corpus specifically for bibliometric research would be very useful.

3. ACL ARC Overview

We describe here the current ACL ARC release[3] and the selection and standardization process used to create it. This current release of the ACL ARC corresponds to the ACL Anthology website as of February 2007, and consists of:

- the source PDF files corresponding to 10,921 articles from the February 2007 snapshot of the Anthology;
- automatically extracted text for all these articles; and
- metadata for the articles, consisting of BibTeX records derived either from the headers of each paper or from metadata taken from the Anthology website.

Whether usable text could be extracted from the document formed the basis of document selection in the ACL ARC. While the Anthology website contained more than these 10,921 PDF sources at the time the corpus was created, those which were problematic in extracting text were excluded.

Automatic text extraction from PDF is known to be problematic (Lawrence et al., 1999). Approaches to text extraction can be categorized as either OCR- or non-OCR-based. Non-OCR approaches try to extract text directly from the PDF data file, whereas OCR approaches use a PDF interpreter to render an image over which standard optical character recognition software is then run to re-capture the text. For the current ACL ARC release, we used PDFBox 0.72[4] to perform direct, non-OCR based text extraction, due to its cost (free), availability and processing speed. This usually resulted in variable quality; results vary from very clean text to completely garbled output. Rather than subjectively selecting a level to threshold extraction results, we included in this corpus release all source PDF articles that produced non-empty output and excluded those generated no output or produced fatal errors in the automated text extraction phase. Excluded documents amounted to 476 papers (about 4% of all available at the time).

Problems that prevent text from being extracted from the PDFs primarily stem from the way font and glyph information is encoded in the source file: when a custom font encoding is used by the PDF generator (often used to make the PDF more compact), the extractable text becomes gibberish.

The metadata consists of the unique ID assigned to each paper, along with the paper's author(s), title, publication venue, and year of publication. The ID is composed of a letter signifying the journal, conference, or workshop series where the paper was presented, the year of publication, and

[3] Version 20080325, available at http://acl-arc.comp.nus.edu.sg/.

[4] http://www.pdfbox.org/.

Total Articles	10,921
Total References	152,546
References to articles inside ACL ARC	38,767 (25.4%)
References to articles outside ACL ARC	113,779 (74.6%)

Table 2: General Statistics for the ACL ARC.

a unique number in that series in that year. So, for example, the ID P00-1004 refers to the fourth paper (1004: paper numbering in a volume starts at 1001) in the 2000 proceedings (represented by the '00') of the main conference of the ACL (represented by the 'P'). The ACL Anthology website includes article metadata for all of the papers in ACL ARC. General statistics for the ACL ARC are shown in Table 2.

However, during the construction of the corpus we found that this metadata was not always correct; the verification of the article metadata revealed errors which have been passed to the Anthology editor for incorporation to the Anthology proper.

The form the metadata takes is as follows: for each series and year, the website provides a list of links to the papers with their associated metadata, i.e.: *P00-1001: Susan E. Brennan. Invited Talk: Processes that Shape Conversation and their Implications for Computational Linguistics.*

The current ACL ARC release specifies the exact identity of the documents in the collection, the documents themselves (in original PDF and converted text versions) and includes gold standard ground truth for document metadata, which allows the evaluation of automated document metadata extraction algorithms that process headers of papers (i.e., title page and abstracts).

4. Future ACL ARC development

The corpus described above is already useful in its own right, by virtue of providing a fixed set of documents that the present consortium of authors has agreed to use for benchmarking. However, we believe that some specific enrichments would make it a useful testbed for an enlarged set of research problems. To enable such research, we have planned for multiple corpus releases that provide data and ground-truth for such research. Future corpus releases will enlarge the corpus with a larger set of documents (as future NLP research publications are archived within the Anthology) and provide both manually validated gold-standard data and automatic processing results of tools run on the corpus. Ground-truth data enables the objective evaluation of OCR benchmarking, information retrieval studies on specific queries and bibliometric research on citation structure; the results of running other tools on the data provides input for more sophisticated processing.

The provision of a standardized collection of documents enables researchers to conduct research on topics of interest to the Digital Libraries and NLP communities. Basic processing such as OCR benchmarks can be run on this corpus, which represents a genre-specific (i.e., academic discourse) corpus. Information retrieval studies may investigate the relevance of research documents given scientific queries. Bibliometric research can analyze the citation structure of this closed collection of documents to programmatically identify key authors and topics in NLP across a span of over 30 years.

Work on the next corpus release focuses on expanding the gold-standard data for both intra-article and inter-article analysis. In particular, we plan to make available ground-truth reference data for the following tasks:

- Intra-article linkages between the sentences containing explicit citations and the reference list items they correspond to[5]. Matching citations to reference items is often straightforward, but determining the scope of the citation—precisely what in the text the citation attaches to—is generally non-trivial. The scope often crosses clausal and sentential boundaries, extending to preceding or subsequent clauses and sentences. Gold-standard data will enable future learning-based methods to address the robustness of work in this area; this research is driven by Macquarie University (Powley and Dale, 2007).

  As an example, consider the citation to **P83-1019** as it appears in example paper P00-1001: ... *Few approaches to parsing have tried to handle disfluent utterances (notable exceptions are Core & Schubert, 1999;* **Hindle, 1983***; Nakatani & Hirschberg, 1994; Shriberg, Bear, & Dowding, 1992).* Ideally, we would like to be able to automatically determine that P83-1019 offers an approach to the parsing of disfluent utterances.

- Inter-article linkage between each reference to its target article, where that article exists in the ACL ARC. By definition, this extends the gold-standard metadata provided for each paper to include the clean metadata for referenced documents. This work is being carried out at the University of Michigan as part of the ACL Anthology Network (AAN) (Joseph and Radev, 2007). Where an article links to another article within the ACL ARC, the ACL ARC identifier is used, otherwise the full reference string is given. This data will be distributed as a text file in the corpus that lists each citation as a separate line, in the form of *citing_paper* ==> *target_paper* . The file is exactly 152,546 lines, corresponding to the total number of citations made from all papers, as given in Table 2.

  As an example, the same paper P00-100 makes a total of 38 references of which 2 point to other papers in the ACL ARC collection: H92-1085 and P83-1019. Such gold-standard data will enable, amongst other goals, the exact construction of the social network of NLP researchers within the ACL ARC, and subsequent visualization and exploration.

[5] Note that we adopt the term 'reference' to refer to bibliographic information found at the end of an article (in the reference list) and 'citation' to refer to an embedded pointer to the respective reference that appears in the body text. These elements are also distinct from the metadata obtainable from the header of the paper, which often contains additional author information, such as email addresses. Work which is primarily concerned with only one or two of these elements tends to use the terms interchangeably; however, since the ACL ARC contains all three types of information, we explicitly differentiate these data sources with a more precise use of terminology.

Other currently planned work includes: (1) the automatic processing of the ACL ARC documents through an OCR-based text extraction process, to be carried out by multiple sites; (2) automated keyphrase extraction (Nguyen and Kan, 2007), (3) presentation-to-article alignment (Kan, 2007); and (4) the automatic segmentation of references into their constituent fields. The latter three tasks are being carried out by the National University of Singapore. In the same vein, we hope the community will contribute more data and processing results to incorporate into future ACL ARC releases.

We plan to release a new version of the corpus every one to two years to ensure that the community has enough time to utilize the resource for comparative research. We believe that more frequent corpus releases would hamper benchmarking and other comparative research.

5. Related Work

Here, we touch upon related problems in bibliographic data processing, and then describe work that will utilize the ACL ARC as a canonical data source to further develop scholarly article processing.

Reference Segmentation: When references are extracted as full strings from the references section of a PDF document, being able to identify the separate data fields that make up these reference strings (e.g., title, author, venue, page numbers and year of publication) helps subsequent processing steps significantly. However, the different styles adopted for formatting references makes segmentation non-trivial. Different disciplines, publishers, or domains tend to have their own unique styles for the formatting of bibliographic data in the references section of a paper; and scholars occasionally invent their own styles by ignoring (inadvertently or not) the specified style. High accuracy reference segmentation is thus a significant challenge, and one that has already attracted considerable attention (see, for example, (Peng and McCallum, 2004)).

Reference–Article Matching: In order to create links between a reference and the target article, one needs to determine whether a reference matches the (header) metadata for a candidate target article. One can view this matching problem as a specialization of the more general Entity Resolution (or Record Linkage) problem common in the database and data mining communities. Scholars have generally exploited domain-specific characteristics to inform the similarity computation. In bibliographic data, approaches include culling evidence from collaboration networks, and viewing references as artifacts of a probabilistic language model, as well as techniques for linking abbreviated forms to full forms (e.g., *John Doe* and *J. Doe*, or *ACL* and *Association for Computational Linguistics*) or data cleaning methods for fixing errors. All such techniques can significantly help the citation matching process (Kan and Tan, 2008; On, 2007).

5.1. Research Enabled by the ACL ARC

Citation Classification: Citations made in articles serve different purposes, providing a foundation for an article's current focus, pointing to tools with which the research was performed or serving as a contrast to the results given by the article. Work on citation classification tries to automatically determine the purpose of a citation; this work hinges on the correct resolution of the citation to the appropriate reference, and learning the function of lexical cues within citing sentences. Work has already been done in this area on corpora in NLP (Teufel et al., 2006) and in the biomedical domain (Schwartz et al., 2007).

Automatic Survey Article Generation: The iOPENER Project (Information Organization for PENning Expositions on Research), a newly-commenced NSF-funded collaboration between the University of Maryland and the University of Michigan, will link automatic summarization (e.g., (Zajic et al., 2007; Radev et al., 2004; Radev et al., 2005)) and visualization work with citation classification. Key developments in this work will include extending techniques in summarization to handle redundancy, contradictions, and temporal ordering based on citation analyses (Elkiss et al., 2008). The intended result is a set of readily-consumable surveys of different scientific domains and topics, targeted to different audiences and levels. The project will leverage existing publicly-available resources such as the ACL Anthology, ACM Digital Library, CiteSeer, and others for analysis, retrieval, selection, and survey/timeline creation and visualization. The iOPENER software and resulting surveys and timelines will be made publicly available.

5.2. Relationship to Grassroots Initiatives

At the Association for Computational Linguistics 2007 conference in Prague, the ACL Executive Committee called for grassroots proposals for activities that would benefit the community. Three proposals centered on developments related to the ACL Anthology, which we refer to here as the Linked Anthology, the Extended Anthology and the Video Archives. The work reported here is an outcome of the Linked Anthology proposal. The Linked Anthology additionally specifies the creation of tools for bibliographic data processing and proposes that any corrected gold-standard data be propagated to the Anthology (e.g., allowing citations in the body of the PDF version of a conference paper to link directly to the target PDF paper). The Extended Anthology and Video Archives aim to extend the reach of Anthology, by including, respectively, grey literature (e.g., institutional technical reports) and multi-modal records (e.g., videos of conference presentations (Lee, 2007)). If and when the Anthology incorporates these additional resources, future releases of the the ACL ARC will, where practically possible, also incorporate these additional corpora.

6. Discussion and Conclusion

Aside from the Anthology, as indicated in the introduction to this paper, quite a few digital anthologies now exist, and some of these far exceed the Anthology in terms of size as well as breadth; in our community, perhaps the most widely known is example is the ACM Digital Library (White, 2001). The skeptic will rightly question why the ACL ARC is a significant reference corpus in light of these other resources. What distinguishes this work is that it is both collaborative and standardized. Several research teams, representing ACL's worldwide membership, have joined to develop the ACL ARC. This collaboration will propose standard tasks (e.g., text extraction and reference segmentation) that can integrate with the community's standard venues for bake-off competitions (e.g., CoNLL). The standardization aspect is possibly more crucial, as live digital anthologies are diachronic, being updated on a daily basis; in contrast, a reference corpus needs to be frozen in order to facilitate comparison. By versioning and publishing only major revisions, we hope that the ACL ARC will facilitate performance comparisons.

While other communities also have digital anthologies, for example DBLP (Ley, 2002), many researchers look towards the NLP community to provide leadership towards the next generation of scholarly digital libraries that will be enhanced by sophisticated language processing. We believe this is a challenge we should address head-on, and what better place to start than a set of data we are already familiar with? The creation of the ACL ARC offers an opportunity to bring together researchers from various disciplines (such as NLP, DB and IR) to research and implement tools and resources which will provide a basis for the future of academic research. We call on the community to become involved in this exciting development, where we can utilize our own technology to advance and highlight our research.

7. Acknowledgments

We would like to acknowledge the support of the ACL Executive Committee in their drive to support the computational linguistics community's efforts in this area. Work at the University of Michigan work was partially supported by the National Science Foundation under grants 0534323 and 0705832. All opinions, findings, conclusions and recommendations in any material resulting from this workshop are those of the participants, and do not necessarily reflect the views of the National Science Foundation. We would also like to thank Nayeoung Kim and Paul Hartzog for their annotation help.

8. References

Aaron Elkiss, Siwei Shen, Anthony Fader, Güneş Erkan, David States, and Dragomir R. Radev. 2008. Blind men and elephants: what do citation summaries tell us about a research article? *Journal of the American Society for Information Science and Technology*, 59(1):51–62.

Mark T. Joseph and Dragomir R. Radev. 2007. Citation analysis, centrality, and the acl anthology. Technical Report CSE-TR-535-07, University of Michigan. Department of Electrical Engineering and Computer Science.

Min-Yen Kan and Yee Fan Tan. 2008. Record matching in digital library metadata. *Communications of the ACM (CACM)*, 51(2):91–94.

Min-Yen Kan. 2007. SlideSeer: A digital library of aligned document and presentation pairs. In *Proceedings of the Joint Conference on Digital Libraries (JCDL '07)*, pages 81–90, Vancouver, Canada, June.

Steve Lawrence, C. Lee Giles, and Kurt Bollacker. 1999. Digital libraries and autonomous citation indexing. *IEEE Computer*, 32(6):67–71.

Dongwon Lee. 2007. Toward web-scale academic video search engine. Technical report, The Pennsylvania State University, December.

Michael Ley. 2002. The DBLP computer science bibliography: Evolution, research issues, perspectives. In *International Symposium on String Processing and Information Retrieval (SPIRE)*, pages 1–10, September.

Thuy Dung Nguyen and Min-Yen Kan. 2007. Keyphrase extraction in scientific publications. In *Proc. of International Conference on Asian Digital Libraries (ICADL '07)*, pages 317–326, Hanoi, Vietnam. Springer.

Byung-Won On. 2007. *Data Cleaning Techniques by means of Entity Resolution*. Ph.D. thesis, The Pennsylvania State University, May.

Fuchun Peng and Andrew McCallum. 2004. Accurate information extraction from research papers using conditional random fields. In *Proceedings of Human Language Technology Conference / North American Chapter of the Association for Computational Linguistics annual meeting*, pages 329–336.

Brett Powley and Robert Dale. 2007. Evidence-based information extraction for high accuracy citation and author name identification. In *Proceedings of RIAO 2007: the 8th Conference on Large-Scale Semantic Access to Content*, Pittsburgh, Pa., USA, 30 May to 1 June.

Dragomir R. Radev, Timothy Allison, Sasha Blair-Goldensohn, John Blitzer, Arda Celebi, Stanko Dimitrov, Elliott Drabek, Ali Hakim, Wai Lam, Danyu Liu, Jahna Otterbacher, Hong Qi, Horacio Saggion, Simone Teufel, Michael Topper, Adam Winkel, and Zhu Zhang. 2004. MEAD: A platform for multidocument multilingual text summarization. In *LREC*, Lisbon, Portugal, May.

Dragomir R. Radev, Jahna Otterbacher, Adam Winkel, and Sasha Blair-Goldensohn. 2005. NewsInEssence: Summarizing online news topics. *Communications of the ACM*, 48(10):95–98.

Ariel Schwartz, Anna Divoli, and Marti Hearst. 2007. Multiple alignment of citation sentences with conditional random fields and posterior decoding. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 847–857.

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, Sydney, Australia, July. Association for Computational Linguistics.

John White. 2001. ACM opens portal to computing literature. *Communications of the ACM (CACM)*, 44(7):14–16,28, July.

David M. Zajic, Bonnie J. Dorr, Jimmy Lin, and Richard Schwartz. 2007. Multi-candidate reduction: Sentence compression as a tool for document summarization tasks. *Information Processing and Management Special Issue on Summarization*, 43(6):1549–1570.