# The ACL Anthology Reference Corpus: a reference dataset for bibliographic research

**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]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; but we believe that it can also be an object of study in its own right. We describe a 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 across the world. Our goal is to make the corpus widely available, and to encourage other researchers to use it for experiments as a standard testbed for 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: document collection, reference metadata extraction and cleaning, and infrastructure for searching and browsing. High performance on these tasks is critical to lightweight, low-cost maintenance of quality 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 to measure performance on these key infrastructural tasks.

The ACL Anthology represents the community's most up-to-date resource on NLP research, in which newly-published conference proceedings and journal articles extend the collection several times a year. 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 comparison across experiments difficult. Research communities such as 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. To facilitate future work, a standardized reference corpus is needed.

This paper describes the ACL Anthology Reference Corpus[1] (ACL ARC), a collaborative atttempt to standardize

| Name | Domains | # Articles | # References | Source |
|---|---|---|---|---|
| ISI SCI | Sciences | 0 | 25 | HH |
| http://portal.isiknowledge.com/portal.cgi | | | | |
| CAS | Chemistry | 0 | 23 | HH |
| http://www.cas.org/ | | | | |
| PubMed | Life Science | 0 | 12 | HH |
| http://www.ncbi.nlm.nih.gov/sites/entrez | | | | |
| CiteSeer | Sciences | 0.8 | 10 | SS |
| http://citeseer.ist.psu.edu/ | | | | |
| arXiv e-Print | Physics, Math | 0.3 | 0.3 | HS |
| http://arxiv.org/ | | | | |
| SPIRES-HEP | High-energy Physics | 0.27 | 0.5 | HH |
| http://www.slac.stanford.edu/spires/index.shtml/hep/ | | | | |
| DBLP | Computer Science | 0 | 0.93 | HH |
| http://www.informatik.uni-trier.de/ ley/db/index.html | | | | |
| CSB | Computer Science | 0 | 2 | SS |
| http://liinwww.ira.uka.de/bibliography/ | | | | |
| ACM DL | Computer Science | N/A | N/A | HS |
| http://portal.acm.org/dl.cfm | | | | |
| IEEE DL | Engineering | N/A | N/A | HS |
| http://www.computer.org/portal/site/csdl/ | | | | |
| SIGMOD Anthology | Computer Science | N/A | N/A | HH |
| http://www.informatik.uni-trier.de/ ley/db/anthology.html | | | | |
| Google Scholar | Sciences | N/A | N/A | SS |
| http://scholar.google.com/ | | | | |

Table 1: Demographics of a sample of current scholarly digital collections. Sizes are in millions of PS or PDF articles held by the collection. Source gives 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.

[†]Contact author

[1]For ease of reference we use "ACL ARC" to refer to the corpus project under discussion, and "ACL Anthology" for the publicly-accessible website (http://aclweb.org/anthology-index/) containing the ACL publication archives, which currently spans

a reference corpus for the Anthology. We first 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. In Section 3., we describe long-term plans for the ACL ARC that include future corpus releases and bibliographic processing the development of baseline tools for scholarly document processing. that represent state-of-the-art algorithms that can be used for comparison. In Section 4., we review related work in bibliographic research and discuss how ACL ARC's development 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. ACL ARC Overview

We describe the current ACL ARC release[2] 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 from the headers of each paper or metadata taken from the Anthology website.

The metadata consists of an ID assigned to each paper, the papers author(s), title, venue, and year. The ID is composed of a letter signifying the journal, conference, or workshop where the paper was presented, the two digit year, and a unique number.

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 are also distinct from the metadata obtainable from the header of the paper (often containing additional author information, such as email addresses). The community often refers to these terms interchangably when handling only one of these information sources, but as the ACL ARC contains all three types of information, we must differentiate these data sources.

While more PDF sources existed on the Anthology website, those which generated no output or produced fatal errors in the automated text extraction phase were excluded from the corpus; these amounted to 476 papers (about 4% of all available at the time). Automatic text extraction from PDF is known to be problematic (Lawrence et al., 1999), and approaches to the task 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 run to recapture the text. For the current ACL ARC release, we have used PDFBox 0.38[3] 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, often due to the way font and glyph information is encoded in the source PDF file. 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.

| | |
|---|---|
| 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 of the ACL ARC.

The ACL Anthology website included the article metadata for all of the papers in ACL ARC that were either manually entered by authors or the Anthology editor. However, during the construction of the corpus we found that these metadata were not always correct – the verification of the article metadata revealed some errors, which have been passed to the management of the Anthology for its website revision.

The form the metadata takes is as follows: for each venue/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).

## 3. Future ACL ARC development

The corpus described above is already useful in its own right by declaring a fixed set of documents that this consortium of authors have 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 NLP research progresses and is archived within the Anthology) and provide both manually validated gold-standard data and automatic processing results of tools run on the corpus. Such ground truth enables the objective evaluation of OCR benchmarking, information retrieval studies on specific queries and bibliometric research on citation structure.

Such 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

the period from the 1970s to 2007.

[2]Version 20071031, available at http://acl-acr.comp.nus.edu.sg/

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

closed collection of documents to programmatically identify key authors and topics in NLP across a span of over 30 years.

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

- Intra-article linkage between the sentences containing explicit citations to the appropriate reference item. Matching citations to reference items is often straightforward, but deciding the scope of the citation within the sentence is non-trivial. The scope often crosses sentential boundaries, extending to subsequent sentences. The gold standard data will enable future learning-based methods to address the robustness of this work. This research is driven by Macquarie University (Powley and Dale, 2007).

  Example (context for **P83-1019** in 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).*

- 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 clear metadata for referenced documents. This research is being done at the University of Michigan, and is described in a separate LREC submission. Such gold-standard data will enable exploration of the social network of NLP scientists, among other goals.

  Example: P00-1001 → P83-1019

Other currently planned work includes: (1) the automatic processing of the ACL ARC documents through a OCR based text extraction process to be done 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 by fields. The latter three tasks are being done by the National University of Singapore. 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. More frequent corpus releases would hamper benchmarking and other comparative research.

## 4. Related Work

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 separate fields of reference strings (e.g., title, author, venue, and year) 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 in formatting citations in the reference section of papers. Scholars invent their own styles by ignoring (inadvertently or not) the specified style. High accuracy reference segmentation is thus a challenge that has been tackled by learning-based graphical NLP methods (Peng and McCallum, 2004).

**Reference-Article Matching.** In order to create links between a reference and the target article, one needs to match if 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 used generally exploited domain-specific characteristics to inform the similarity computation. In bibliographic data, approaches include culling evidence from collaboration networks, viewing references as artifacts of a probabilistic language model, as well as 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 typographical errors can significantly help the success of the citation matching process (Kan and Tan, 2008).

### 4.1. Research enabled by 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. This work hinges on the correct resolution of the citation to the appropriate reference and learning the function of lexical cues within citation sentences. Work has already been done 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), an NSF-funded collaboration between the University of Maryland and the University of Michigan, which has just started, will link automatic summarization (e.g., (Zajic et al., 2007; Radev et al., 2004)) 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., 2007). 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.

### 4.2. 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 the ACL Anthology: 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 suggests 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). Both the Extended Anthology and Video Archives depend on extending the reach of Anthology, to include grey literature (e.g., institutional technical reports) and multi-modal records (e.g., videos of conference presentations), respectively. If and when the Anthology incorporates these additional resources, future releases of the the ACL ARC will incorporate these additional corpora as well, where practically possible.

## 5. Discussion and Conclusion

The ACL Anthology has been one of the natural language processing (NLP) community's longest-standing resources of freely accessible research. Steven Bird proposed the initiative to the ACL Executive at the 2001 ACL conference, in response to a call for something to mark the ACL's 40th anniversary. In the following 12 months, over US$50,000 of institutional and individual was donated funding efforts to digitize all previous two decades of ACL conference and journal issues. Pages were scanned at 600dpi grayscale for archival storage, and then down-sampled 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 OCR 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, 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.

Currently, the ACL's conference publication software automatically generates conference proceedings that can be incorporated into the Anthology with a minimum of manual effort. At the time of writing, the Anthology contains 14,000 articles, indexed by a host of other digital libraries and repositories, such as Citeseer, Google Scholar, OLAC, and the ACM Digital Library.

Aside from the Anthology, quite a few digital anthologies now exist – e.g., ACM Digital Library (White, 2001) – that far exceed the Anthology in terms of size as well as breadth. 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, reference segmentation) that can integrate with the community's standard venues for bakeoff 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 frozen, 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. We believe it is a challenge that is both possible and practical. The creation of the ACL ARC will bring researchers together from various disciplines (such as NLP, DB and IR) to research and implement 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.

## 6. Acknowledgments

We would like to acknowledge the support of the ACL Executive Committee in their drive to support the computational linguistics community's efforts.

## 7. References

Aaron Elkiss, Siwei Shen, Anthony Fader, David States, and Dragomir Radev. 2007. Blind men and elephants: What do citation summaries tell us about a research article? *Journal of the American Society for Information Science*, January. To be submitted.

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

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)*, Vancouver, Canada, June.

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

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)*, Hanoi, Vietnam, December. To appear.

Fuchun Peng and Andrew McCallum. 2004. Accurate information extraction from research papers using conditional random fields. In *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. *Recherche d'Information Assist*.

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.

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, June.

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*. To appear.