The ACL Anthology Network Corpus as a Resource for NLP-based Bibliometrics

Amjad Abu-Jbara AMJBARA@UMICH.EDU
EECS Department, University of Michigan, Ann Arbor, MI 48105 USA

Dragomir R. Radev RADEV@UMICH.EDU
EECS Department and School of Information, University of Michigan, Ann Arbor, MI 48105 USA

The ACL Anthology[1] is one of the most successful initiatives of the Association for Computational Linguistics (ACL). It was initiated by Steven Bird in 2001 and is now maintained by Min-Yen Kan. It includes all papers (20,000+) published by ACL and related organizations as well as the Computational Linguistics journal over a period of four decades.

The ACL Anthology Network (AAN)[2] is another successful project built on top of the ACL Anthology. It was started in 2007 by our group (CLAIR) (Radev et al., 2009) at the University of Michigan. Table 1 shows some statistics of the current release of AAN. We convert the articles included in the ACL Anthology corpus (excluding book reviews) from PDF to text. This text is then processed to identify tokens, mark sentence boundaries, identify the abstracts, extract key terms, remove end-of-line hyphens, identify section boundaries and section headings, identify references, etc. We also extract names from paper metadata, normalize them, and match them to existing authors. Moreover, we extract author-paper affiliations and manually classify authors genders.

AAN provides manually extracted citation and collaboration networks of the ACL anthology articles and their authors. It also includes rankings of papers and authors based on their centrality statistics in the citation and collaboration networks. It also includes the citing sentences associated with each citation link. These sentences were extracted automatically and then cleaned manually. The fact that AAN is manually curated and annotated makes it useful to evaluate various NLP components.

We also developed methods to automatically identify non-explicit citing sentences (Qazvinian & Radev, 2010) (i.e. sentences that talk about the cited article but do *not* contain an explicit reference to it) and to identify the scope of a given reference in citing sentences that contain multiple references (Abu Jbara & Radev, 2012).

Number of papers	19,647
Number of authors	16,152
Number of venues	356
Number of paper citations	94,973
Citation network diameter	22
Collaboration network diameter	15

Table 1. Statistics of AAN 2012 release

In addition, we developed supervised methods that use the explicit and non-explicit citing sentences to identify the purpose and sentiment of citation (Abu-Jbara et al., 2013). By citation purpose, we mean the author's intention behind selecting the cited paper and commenting on it. We recognize five citation purposes including *Critique*, *Comparison*, *Use*, *Substantiation*, *Basis*, and *Other*. Citation sentiment refers to whether the cited article is positively, negatively, or neutrally evaluated by the author of the citing paper.

We also developed methods for generating summaries of a scientific article using the explicit and non-explicit citing sentences that refer to it (Qazvinian & Radev, 2008; Qazvinian et al., 2010; Abu-Jbara & Radev, 2011). We extended this work to generate surveys of scientific topics by selecting a set of papers related to the topic of interest and then summarizing them using their citing sentences as the source text (Mohammad et al., 2009; Qazvinian et al., 2013; Jha et al., 2013).

In recent work, we used AAN to investigate the problem of predicting the future impact of publications. We base the predictions on various types of indicators including author impact history, venue impact, content

[1] http://www.aclweb.org/anthology-new/

[2] http://aan.eecs.umich.edu

analysis, purpose and sentiment analysis of outgoing citations, etc.

AAN has also been used in several other studies such as citation prediction (Yogatama et al., 2011), survey generation (Dunne et al., 2012), academic collaboration modeling (Johri et al., 2011), analyzing research dynamics (Gupta & Manning, 2011), stylometric analysis of scientific articles (Bergsma et al., 2012), publication ranking (Jiang et al., 2012), author gender analysis (Vogel & Jurafsky, 2012), science networks analysis (Gove et al., 2011; Wu et al., 2010), research factions identification (Sim et al., 2012), document clustering (Muthukrishnan et al., 2011), and link prediction (Chaturvedi et al., 2012).

AAN and the aforementioned uses and applications that utilize it can be employed to aid the peer reviewing processes of scientific publications. In the rest of this paper, we present some examples and ideas.

- **Suggesting reviewers and related papers.** The techniques that we developed to automatically select related papers for the survey generation task can be used in the review process to help reviewers quickly identify the important previous work related to the paper under review. Similarly, area chairs can use the same techniques to identify potential reviewers among those who published work related to the area of the reviewed paper.

- **Summarization and survey generation.** For reviewers who are not completely familiar with the topic of the reviewed paper, our methods for survey generation and scientific article summarization can be used to provide reviewers with automatic customized surveys and summaries about the topic if such surveys do not exist.

- **Automatic sentiment analysis of reviews.** Reviews can be treated as citations. Each review is a critical summary of the reviewed paper that evaluates its merits and/or faults from the viewpoint of the reviewer. Applying the techniques we developed for analyzing the sentiment of citations to reviews can help area chairs to quickly recognize the strengths and the weaknesses that each reviewer identified in the reviewed paper. This analysis can be also used to double check that the scores that the reviewers gave to the paper are consistent with the comments they wrote about it.

- **Automatic resolution of reviews disagreement**. When two reviews provide inconsistent or contradicting evaluation (scores) of the reviewed paper, the area chair needs to resolve this disagreement usually by asking the reviewers to discuss the points of disagreement. This process can be expedited if the points of disagreement can be automatically identified through analyzing the text of reviews.

- **Extraction of methods, tools, etc.** Term extraction and term classification techniques can be applied to the reviewed article to highlight the important contributions of the paper (new method, data sets, tools, etc).

- **Automatic review generation.** Several of the tasks that the reviewers often do can be automated or facilitated through the aid of automatic methods. For example, the quality of writing (spelling, grammaticality, lexical choice, etc.) can be evaluated automatically. For another example, context-aware citation recommendation techniques (He et al., 2010; Tang & Zhang, 2009; He et al., 2011) can be used to suggest citations to authors or to detect claims that are missing supporting citations.

- **Rhetorical-level analysis of reviewed articles.** Argumentative zoning methods (Teufel & Moens, 2000; Guo et al., 2011) can be used to help the reviewers quickly find pieces of text in the reviewed article that provide background information about the topic, present the goal of the paper, present the contribution of the authors, compares current work to previous work, etc.

References

Abu-Jbara, Amjad and Radev, Dragomir. Coherent citation-based summarization of scientific papers. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 500–509, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P11-1051.

Abu Jbara, Amjad and Radev, Dragomir. Reference scope identification in citing sentences. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 80–90, Montréal, Canada, June 2012. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/N12-1009.

Abu-Jbara, Amjad, Ezra, Jefferson, and Radev, Dragomir R. Purpose and polarity of citation: Towards nlp-based bibliometrics. In *Proceedings of* *the North American Association for Computational Linguistics*, 2013.

Bergsma, Shane, Post, Matt, and Yarowsky, David. Stylometric analysis of scientific articles. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pp. 327–337, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics. ISBN 978-1-937284-20-6. URL http://dl.acm.org/citation.cfm?id=2382029.2382071.

Chaturvedi, Snigdha, Daumé III, Hal, Moon, Taesun, and Srivastava, Shashank. A topical graph kernel for link prediction in labeled graphs. In *Proceedings of the Workshop on Mining and Learning with Graphs (MLG-2012)*, 2012.

Dunne, Cody, Shneiderman, Ben, Gove, Robert, Klavans, Judith, and Dorr, Bonnie. Rapid understanding of scientific paper collections: Integrating statistics, text analytics, and visualization. *Journal of the American Society for Information Science and Technology*, 63(12):2351–2369, 2012.

Gove, Robert, Dunne, Cody, Shneiderman, Ben, Klavans, Judith, and Dorr, Bonnie. Understanding scientific literature networks: An evaluation of action science explorer. Technical report, Citeseer, 2011.

Guo, Yufan, Korhonen, Anna, and Poibeau, Thierry. A weakly-supervised approach to argumentative zoning of scientific documents. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 273–283, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/D11-1025.

Gupta, Sonal and Manning, Christopher. Analyzing the dynamics of research by extracting key aspects of scientific papers. In *Proceedings of 5th international joint conference on natural language processing*, pp. 1–9, 2011.

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

He, Qi, Kifer, Daniel, Pei, Jian, Mitra, Prasenjit, and Giles, C Lee. Citation recommendation without author supervision. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pp. 755–764. ACM, 2011.

Jha, Rahul, Abu-Jbara, Amjad, and Radev, Dragomir R. A system for summarizing scientific topics starting from keywords. In *Proceedings of The Association for Computational Linguistics (short paper)*, 2013.

Jiang, Xiaorui, Sun, Xiaoping, and Zhuge, Hai. Towards an effective and unbiased ranking of scientific literature through mutual reinforcement. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, CIKM '12, pp. 714–723, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1156-4. doi: 10.1145/2396761.2396853. URL http://doi.acm.org/10.1145/2396761.2396853.

Johri, Nikhil, Ramage, Daniel, McFarland, Daniel A., and Jurafsky, Daniel. A study of academic collaboration in computational linguistics with latent mixtures of authors. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, LaTeCH '11, pp. 124–132, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. ISBN 9781937284046. URL http://dl.acm.org/citation.cfm?id=2107636.2107652.

Mohammad, Saif, Dorr, Bonnie, Egan, Melissa, Hassan, Ahmed, Muthukrishan, Pradeep, Qazvinian, Vahed, Radev, Dragomir, and Zajic, David. Using citations to generate surveys of scientific paradigms. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 584–592, Boulder, Colorado, June 2009. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/N/N09/N09-1066.

Muthukrishnan, Pradeep, Radev, Dragomir, and Mei, Qiaozhu. Simultaneous similarity learning and feature-weight learning for document clustering. *Proceedings of textgraphs-6: Graph-based methods for natural language processing*, pp. 42–50, 2011.

Qazvinian, Vahed and Radev, Dragomir R. Scientific paper summarization using citation summary networks. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pp. 689–696, Manchester, UK, August 2008. Coling 2008 Organizing Committee. URL http://www.aclweb.org/anthology/C08-1087.

Qazvinian, Vahed and Radev, Dragomir R. Identifying non-explicit citing sentences for citation-based summarization. In *Proceedings of the 48th Annual*

Meeting of the Association for Computational Linguistics, pp. 555–564, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P10-1057.

Qazvinian, Vahed, Radev, Dragomir R., and Ozgur, Arzucan. Citation summarization through keyphrase extraction. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pp. 895–903, Beijing, China, August 2010. Coling 2010 Organizing Committee. URL http://www.aclweb.org/anthology/C10-1101.

Qazvinian, Vahed, Radev, Dragomir R., Mohammad, Saif, Dorr, Bonnie, Zajic, David, Whidby, Michael, and Moon, Taesun. Generating extractive summaries of scientific paradigms. *Journal of Artificial Intelligence Research (JAIR)*, 2013.

Radev, Dragomir R., Muthukrishnan, Pradeep, and Qazvinian, Vahed. The acl anthology network corpus. In *NLPIR4DL '09: Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, pp. 54–61, Morristown, NJ, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-58-9.

Sim, Yanchuan, Smith, Noah A., and Smith, David A. Discovering factions in the computational linguistics community. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, ACL '12, pp. 22–32, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=2390507.2390511.

Tang, Jie and Zhang, Jing. A discriminative approach to topic-based citation recommendation. In *Advances in Knowledge Discovery and Data Mining*, pp. 572–579. Springer, 2009.

Teufel, Simone and Moens, Marc. What's yours and what's mine: Determining intellectual attribution in scientific text. In *2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pp. 9–17, Hong Kong, China, October 2000. Association for Computational Linguistics. doi: 10.3115/1117794.1117796. URL http://www.aclweb.org/anthology/W00-1302.

Vogel, Adam and Jurafsky, Dan. He said, she said: gender in the acl anthology. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, ACL '12, pp. 33–41, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=2390507.2390512.

Wu, Hao, He, Jun, Pei, Yijian, and Long, Xin. Finding research community in collaboration network with expertise profiling. In *Advanced Intelligent Computing Theories and Applications*, pp. 337–344. Springer, 2010.

Yogatama, Dani, Heilman, Michael, O'Connor, Brendan, Dyer, Chris, Routledge, Bryan R, and Smith, Noah A. Predicting a scientific community's response to an article. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 594–604. Association for Computational Linguistics, 2011.