Lang Resources & Evaluation
DOI 10.1007/s10579-012-9211-2

ORIGINAL PAPER

The ACL anthology network corpus

Dragomir R. Radev · Pradeep Muthukrishnan · Vahed Qazvinian · Amjad Abu-Jbara

Abstract We introduce the ACL Anthology Network (AAN), a comprehensive manually curated networked database of citations, collaborations, and summaries in the field of Computational Linguistics. We also present a number of statistics about the network including the most cited authors, the most central collaborators, as well as network statistics about the paper citation, author citation, and author collaboration networks.

Keywords ACL Anthology Network · Bibliometrics · Scientometrics · Citation analysis · Citation summaries

1 Introduction

The ACL Anthology[1] is one of the most successful initiatives of the Association for Computational Linguistics (ACL). The ACL is a society for people working on problems involving natural language and computation. It was initiated by Steven Bird (2008) and is now maintained by Min Yen Kan. It includes all papers published by ACL and related organizations as well as the Computational Linguistics journal over a period of four decades.

ACL Anthology has a major limitation in that it is just a collection of papers. It does not include any citation information or any statistics about the productivity of the various researchers who contributed papers to it. We embarked on an ambitious initiative to manually annotate the entire Anthology and curate the ACL Anthology Network (AAN).[2]

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

[2] http://clair.si.umich.edu/anthology/.

D. R. Radev · P. Muthukrishnan · V. Qazvinian (✉) · A. Abu-Jbara
Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, USA
e-mail: vahed@umich.edu

Published online: 06 January 2013

Table 1 Statistics of AAN 2011 release

Number of papers	18,290
Number of authors	14,799
Number of venues	341
Number of paper citations	84,237
Citation network diameter	22
Collaboration network diameter	15
Number of citing sentences	77,753

AAN was started in 2007 by our group at the University of Michigan (Radev et al. 2009a, b). AAN provides citation and collaboration networks of the articles included in the ACL Anthology (excluding book reviews). AAN also includes rankings of papers and authors based on their centrality statistics in the citation and collaboration networks, as well as the citing sentences associated with each citation link. These sentences were extracted automatically using pattern matching and then cleaned manually. Table 1 shows some statistics of the current release of AAN.

In addition to the aforementioned annotations, we also annotated each paper by its institution in the goal of creating multiple gold standard data sets for training automated systems for performing tasks like summarization, classification, topic modeling, etc.

Citation annotations in AAN provide a useful resource for evaluations multiple tasks in Natural Language Processing. The text surrounding citations in scientific publications has been studied and used in previous work. Nanba and Okumura (1999) used the term citing area to refer to citing sentences. They define the *citing area* as the succession of sentences that appear around the location of a given reference in a scientific paper and have connection to it. They proposed a rule-based algorithm to identify the citing area of a given reference. In Nanba et al. (2000) they use their citing area identification algorithm to identify the purpose of citation (i.e. the author's reason for citing a given paper). In a similar work, Nakov et al. (2004) use the term *citances* to refer to citing sentences. They explored several different uses of citances including the creation of training and testing data for semantic analysis, synonym set creation, database curation, summarization, and information retrieval.

Other previous studies have used citing sentences in various applications such as: scientific paper summarization (Elkiss et al. 2008; Qazvinian and Radev 2008, 2010; Mei and Zhai 2008; Qazvinian et al. 2010; Abu-Jbara and Radev 2011a), automatic survey generation (Nanba et al. 2000; Mohammad et al. 2009), and citation function classification (Nanba et al. 2000; Teufel et al. 2006; Siddharthan and Teufel 2007; Teufel 2007).

Other services that are built more recently on top of the ACL Anthology include the ACL Anthology Searchbench and Saffron. The ACL Anthology Searchbench (AAS) (Schäfer et al. 2011) is a Web-based application for structured search in ACL Anthology. AAS provides semantic, full text, and bibliographic search in the papers included in the ACL Anthology corpus. The goal of the Searchbench is both to serve as a showcase for using NLP for text search, and to provide a useful tool for

researchers in Computational Linguistics. However, unlike AAN, AAS does not provide different statistics based on citation networks, author citation and collaboration networks, and content-based lexical networks.

Saffron[3] provides insights to a research community or organization by automatically analyzing the content of its publications. The analysis is aimed at identifying the main topics of investigation and the experts associated with these topics within the community. The current version of Saffron provides analysis for ACL and LREC publications as well as other IR and Semantic Web publication libraries.

2 Curation

The ACL Anthology includes 18,290 papers (excluding book reviews and posters). We converted each of the papers from PDF to text using a PDF-to-text conversion tool (www.pdfbox.org). After this conversion, we extracted the references semi-automatically using string matching. The conversion process outputs all the references as a single block of continuous running text without any delimiters between references. Therefore, we manually inserted line breaks between references. These references were then manually matched to other papers in the ACL Anthology using a "k-best" (with $k = 5$) string matching algorithm built into a CGI interface. A snapshot of this interface is shown in Fig. 1. The matched references were stored together to produce the citation network. If the cited paper is not found in AAN, we have 5 different options the user can choose from. The first option is "Possibly in the anthology but not found," which is used if the string similarity measure failed to match the citation to the paper in AAN. The second option, "Likely in another anthology," is used if the citation is for a paper in a related conference. We considered the following conferences as related conferences AAAI, AMIA, ECAI, IWCS, TREC, ECML, ICML, NIPS, IJCAI, ICASSP, ECIR, SIGCHI, ICWSM, EUROSPEECH, MT, TMI, CIKM and WWW.

The third option is used if the cited paper is a journal paper, a technical report, PhD thesis or a book. The last two options are used if the reference is not readable because of an error in the PDF to text conversion or if it is not a reference. We only use references to papers within AAN while computing various statistics.

In order to fix the issue of wrong author names and multiple author identities we had to perform some manual post-processing. The first names and the last names were swapped for a lot of authors. For example, the author name "Caroline Brun" was present as "Brun Caroline" in some of her papers. Another big source of error was the exclusion of middle names or initials in a number of papers. For example, Julia Hirschberg had two identities as "Julia Hirschberg" and "Julia B. Hirschberg." Other numerous spelling mistakes existed. For instance, "Madeleine Bates" was misspelled as "Medeleine Bates." There were about 1,000 such errors that we had to correct manually. In some cases, the wrong author name was included in the metadata and we had to manually prune such author names. For example, "Sofia Bulgaria" and "Thomas J. Watson" were incorrectly included as author names. Also, there were

[3] http://saffron.deri.ie/.

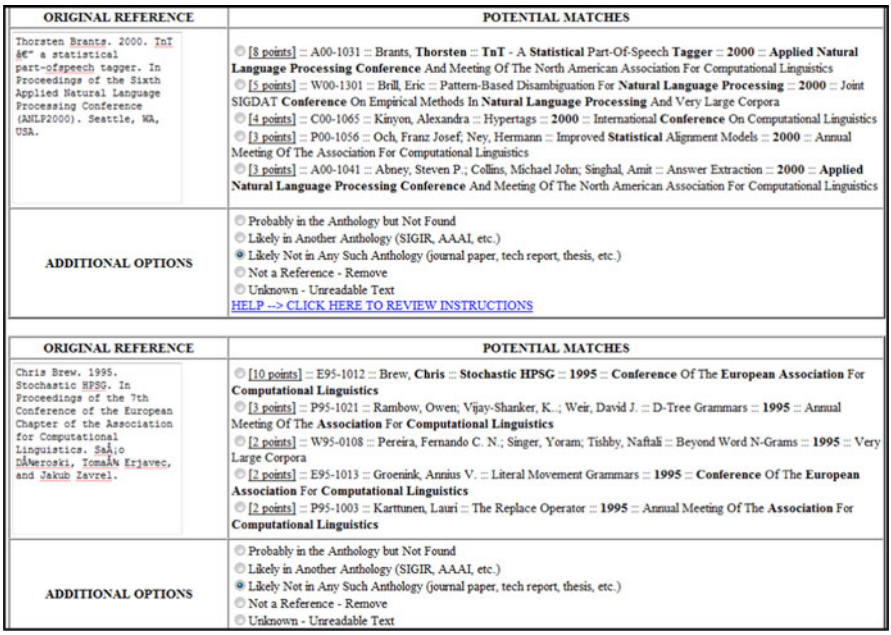

Fig. 1 CGI interface used for matching new references to existing papers

cases of duplicate papers being included in the anthology. For example, C90-3090 and C90-3091 are duplicate papers and we had to remove such papers. Finally, many papers included incorrect titles in their citation sections. Some used the wrong years and/or venues as well. For example, the following is a reference to a paper with the wrong venue.

> Hiroshi Kanayama Tetsuya Nasukawa. 2006. Fully Automatic Lexicon Expansion for Domain-oriented Sentiment Analysis. In ACL.

The cited paper itself was published in EMNLP 2006 and not ACL 2006 as shown in the reference. In some cases, the wrong conference name was included in the metadata itself. For example, W07-2202 had "IJCNLP" as the conference name in the metadata while the right conference name is "ACL". Also, we had to normalize conference names. For example, joint conferences like "COLING-ACL" had "ACL-COLING" as the conference name in some papers.

Our curation of ACL Anthology Networks allows us to maintain various statistics about individual authors and papers within the Computational Linguistics community. Figures 2 and 3 illustrate snapshots of the different statistics computed for an author and a paper respectively. For each author, AAN includes number of papers, collaborators, author and paper citations, and known affiliations as well as h-index, citations over time, and collaboration graph. Moreover, AAN includes paper metadata such as title, venue, session, year, authors, incoming and outgoing citations, citing sentences, keywords, bibtex item and so forth.

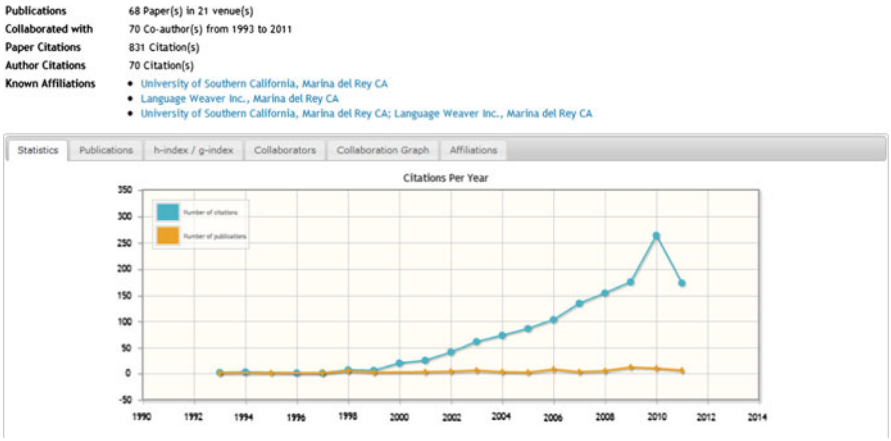

Fig. 2 Snapshot of the different statistics computed for an author

In addition to citation annotations, we have manually annotated the gender of most authors in AAN using the name of the author. If the gender cannot be identified without any ambiguity using the name of the author, we resorted to finding the homepage of the author. We have been able to annotate 8,578 authors this way: 6,396 male and 2,182 female.

The annotations in AAN enable us to extract a subset of ACL-related papers to create a self-contained dataset. For instance, one could use the venue annotation of AAN papers and generate a new self-contained anthology of articles published in BioNLP workshops.

3 Networks

Using the metadata and the citations extracted after curation, we have built three different networks. The paper citation network is a directed network in which each node represents a paper labeled with an ACL ID number and edges represent citations between papers. The paper citation network consists of 18,290 papers (nodes) and 84,237 citations (edges).

The author citation network and the author collaboration network are additional networks derived from the paper citation network. In both of these networks a node is created for each unique author. In the author citation network an edge is an occurrence of an author citing another author. For example, if a paper written by Franz Josef Och cites a paper written by Joshua Goodman, then an edge is created between Franz Josef Och and Joshua Goodman. Self-citations cause self-loops in the author citation network. The author citation network consists of 14,799 unique authors and 573,551 edges. Since the same author may cite another author in several papers, the network may consist of duplicate edges. The author citation network consists of 325,195 edges if duplicates are removed.

In the author collaboration network, an edge is created for each collaborator pair. For example, if a paper is written by Franz Josef Och and Hermann Ney, then an

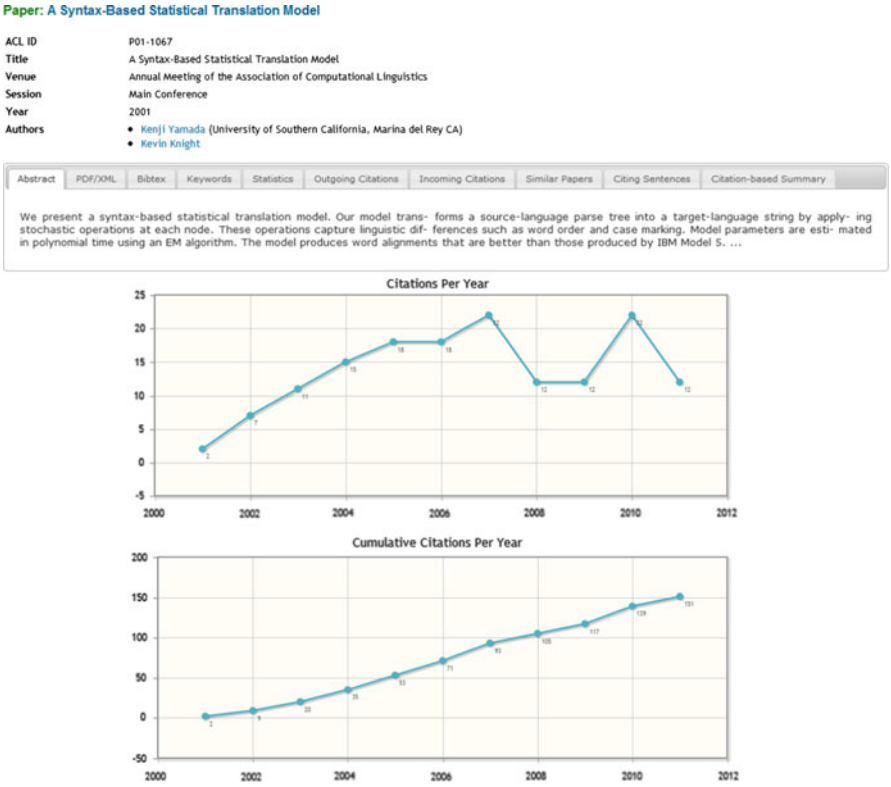

Fig. 3 Snapshot of the different statistics computed for a paper

edge is created between the two authors. Table 2 shows some brief statistics about the different releases of the data set (2008–2011). Table 3 shows statistics about the number of papers in some of the renowned conferences in Natural Language Processing.

Various statistics have been computed based on the data set released in 2007 by Radev et al. (2009a, b). These statistics include modified PageRank scores, which eliminate PageRank's inherent bias towards older papers by normalizing the score by age (Radev et al. 2009a, b), Impact factor, correlations between different measures of impact like h-index, total number of incoming citations, and PageRank. We also report results from a regression analysis using h-index scores from different sources (AAN, Google Scholar) in an attempt to identify multi-disciplinary authors.

4 Ranking

This section shows some of the rankings that were computed using AAN. Table 4 lists the 10 most cited papers in AAN along with their number of citations in Google Scholar as of June 2012. The difference in size of the two sites explains the

Table 2 Growth of citation volume

Years	Network			
		Paper citation network	Author citation network	Author collaboration network
2008	*n*	13,706	11,337	11,337
	m	54,538	196,505	39,963
2009	*n*	14,912	12,499	12,499
	m	61,527	230,658	45,429
2010	*n*	16,857	14,733	14,733
	m	72,463	477,124	52,036
2011	*n*	18,290	14,799	14,799
	m	84,237	573,551	56,966

n number of nodes; *m* number of edges

difference in absolute numbers of citations. The relative order is roughly the same except for the more interdisciplinary papers (such as the paper on the structure of discourse), which are disproportionately getting fewer citations in AAN.

The highest cited paper is (Marcus et al. 1993) with 775 citations within AAN. The next papers are about Machine Translation, Maximum Entropy approaches, and Dependency Parsing. Table 5 shows the same ranking (number of incoming citations) for authors. In this table, the values in parentheses exclude self-citations. Other ranking statistics in AAN include author h-index and authors with the least Average Shortest Path (ASP) length in the author collaboration network. Tables 6, 7 show top 10 authors according these two statistics respectively.

4.1 PageRank scores

AAN also includes PageRank scores for papers. It must be noted that the PageRank scores should be interpreted carefully because of the lack of citations outside AAN. Specifically, out of the 155,858 total number of citations, only 84,237 are within AAN. Table 8 shows AAN papers with the highest PageRank per year scores (PR).

5 Related phrases

We have also computed the related phrases for every author using the text from the papers they have authored, using the simple TF-IDF scoring scheme. Table 9 shows an example where top related words for the author Franz Josef Och are listed.

6 Citation summaries

The citation summary of a paper, *P*, is the set of sentences that appear in the literature and cite *P*. These sentences usually mention at least one of the cited paper's contributions. We use AAN to extract the citation summaries of all articles,

Table 3 Statistics for popular venues

Venue	Number of papers	Number of citations
COLING	3,644	12,856
ACL	3,363	25,499
Computational linguistics	699	12,080
EACL	704	2,657
EMNLP	1,084	7,903
CoNLL	533	3,602
ANLP	334	2,773

Table 4 Papers with the most incoming citations in AAN and their number of citations in Google Scholar as of June 2012

Rank	Citations		Title
	AAN	Google scholar	
1	775	3,936	Building A Large Annotated Corpus Of English: The Penn Treebank
2	615	2,995	The Mathematics Of Statistical Machine Translation: Parameter Estimation
3	591	3,145	Bleu: A Method For Automatic Evaluation Of Machine Translation
4	475	1,408	Minimum Error Rate Training In Statistical Machine Translation
5	473	1,877	A Systematic Comparison Of Various Statistical Alignment Models
6	436	1,711	Statistical Phrase-Based Translation
7	344	1,346	A Maximum Entropy Approach To Natural Language Processing
8	343	2,929	Attention Intentions And The Structure Of Discourse
9	339	1,488	A Maximum-Entropy-Inspired Parser
10	325	1,399	Moses: Open Source Toolkit for Statistical Machine Translation

and thus the citation summary of P is a self-contained set and only includes the citing sentences that appear in AAN papers. Extraction is performed automatically using string-based heuristics by matching the citation pattern, author names and publication year within the sentences.

The example in Table 10 shows part of the citation summary extracted for Eisner's famous parsing paper.[4] In each of the 4 citing sentences in Table 10 the mentioned contribution of (Eisner 1996) is underlined. These contributions are "cubic parsing algorithm" and "bottom-up-span algorithm" and "edge factorization of trees." This example suggests that different authors who cite a particular paper may discuss different contributions (factoids) of that paper. Figure 4 shows a snapshot of the citation summary for a paper in AAN. The first field in AAN citation summaries is the ACL id of the citing paper. The second field is the number of the citation sentence. The third field represents the line number of the reference in the citing paper.

[4] Eisner (1996).

Table 5 Authors with most incoming citations

Rank	Citations	Author name
1 (1)	7,553 (7,463)	Och, Franz Josef
2 (2)	5,712 (5,469)	Ney, Hermann
3 (3)	4,792 (4,668)	Koehn, Philipp
4 (5)	3,991 (3,932)	Marcu, Daniel
5 (4)	3,978 (3,960)	Della Pietra, Vincent J.
6 (7)	3,915 (3,803)	Manning, Christopher D.
7 (6)	3,909 (3,842)	Collins, Michael John
8 (8)	3,821 (3,682)	Klein, Dan
9 (9)	3,799 (3,666)	Knight, Kevin
10 (10)	3,549 (3,532)	Della Pietra, Stephen A.

The values in parentheses are using non-self-citations

Table 6 Authors with the highest h-index in AAN

Rank	h-index	Author name
1	21	Knight, Kevin
2	19	Klein, Dan
2	19	Manning, Christopher D.
4	18	Marcu, Daniel
4	18	Och, Franz Josef
6	17	Church, Kenneth Ward
6	17	Collins, Michael John
6	17	Ney, Hermann

Table 7 Authors with the smallest Average Shortest Path (ASP) length in the author collaboration network

Rank	ASP	Author name
1	2.977	Hovy, Eduard H.
2	2.989	Palmer, Martha Stone
3	3.011	Rambow, Owen
4	3.033	Marcus, Mitchell P.
5	3.041	Levin, Lori S.
6	3.052	Isahara, Hitoshi
7	3.055	Flickinger, Daniel P.
8	3.071	Klavans, Judith L.
9	3.073	Radev, Dragomir R.
10	3.077	Grishman, Ralph

The citation text that we have extracted for each paper is a good resource to generate summaries of the contributions of that paper. In previous work, (Qazvinian and Radev 2008), we used citation sentences and employed a network-based clustering algorithm to summaries of individual papers and more general scientific topics, such as Dependency Parsing, and Machine Translation (Radev et al. 2009a, b).

Table 8 Papers with the highest PageRank per year scores (PR)

Rank	PR	Title
1	955.73	A Stochastic Parts Program And Noun Phrase Parser For Unrestricted Text
2	820.69	Finding Clauses In Unrestricted Text By Finitary And Stochastic Methods
3	500.56	A Stochastic Approach To Parsing
4	465.52	A Statistical Approach To Machine Translation
5	345.11	Building A Large Annotated Corpus Of English: The Penn Treebank
7	318.76	The Contribution Of Parsing To Prosodic Phrasing In An Experimental Text-to-speech system
6	304.11	The Mathematics Of Statistical Machine Translation: Parameter Estimation
8	265.44	Attention Intentions And The Structure Of Discourse
9	194.06	A Maximum Entropy Approach To Natural Language Processing
10	171.25	Word-Sense Disambiguation Using Statistical Methods

Table 9 Snapshot of the related words for Franz Josef Och

	Word	TF-IDF
1	Alignment	3060.29
2	Translation	1609.64
3	Bleu	1270.66
4	Rouge	1131.61
5	Och	1070.26
6	Ney	1032.93
7	Alignments	938.65
8	Translations	779.36
9	Prime	606.57
10	Training	562.10

7 Experiments

This corpus has already been used in a variety of experiments (Qazvinian and Radev 2008; Hall et al. 2008; Councill et al. 2008; Qazvinian et al. 2010). In this section, we describe some NLP tasks that can benefit from this data set.

7.1 Reference extraction

After converting a publication's text from PDF to text format, we need to extract the references to build the citation graph. Up till the 2008 release of AAN, we did this process manually. Table 11 shows a reference string in the text format consisting of 5 references spanning multiple lines.

The task is to split the reference string into individual references. Till now, this process has been done manually and we have processed 155,858 citations of which

Table 10 Sample citation summary of Collins (1996)

In the context of DPs, *this edge based factorization* method was proposed by Eisner (1996)
Eisner (1996) gave a generative model with a *cubic parsing algorithm* based on an *edge factorization of trees*
Eisner (1996) proposed an $O(n^3)$ parsing algorithm for PDG
If the parse has to be projective, Eisner's *bottom-up-span algorithm* (Eisner 1996) can be used for the search

Search:

Source Paper	year	Line	Sentence
A00-2011	2000	10	Many corpus-based MT systems require parallel corpora (Brown et al. , 1990; Brown et al. , 1991; Gale and Church, 1991; **Resnik, 1999**).
C10-2054	2010	16	**Resnik (1999)** mined comparable corpora on the assumption that the pages which are comparable of each other share a similar structure (headers, paragraphs, etc.) when text is presented in many languages in the Web.
E06-1033	2006	6	For languages with limited electronic resources, i.e. low-density languages, however, we cannot use automated techniques based on parallel corpora (Gale and Church, 1991; Melamed, 2000; **Resnik, 1999**; Utsuro et al. , 2002), comparable corpora (Fung and Yee, 1998), or multilingual thesauri (Vossen, 1998).
H05-1069	2005	27	First, parallel corpora, especially accurately aligned parallel corpora are rare, although attempts have been made to mine them from the Web (**Resnik, 1999**).
I05-4010	2005	11	The Web is being explored not only as a super corpus for NLP and linguistic research (Kilgarriff and Grefenstette, 2003) but also, more importantly to MT research, as a treasure for mining bitexts of various language pairs (**Resnik, 1999**; Chen and Nie, 2000; Nie and Cai, 2001; Nie and Chen, 2002; Resnik and Smith, 2003; Way and Gough, 2003).
J03-3001	2003	49	**Philip Resnik (1999)** showed that parallel corpora untit then a promising research avenue but largely constrained to the English-French Canadian Hansardcould be found on the Web: We can grow our own parallel corpus using the many Web pages that exist in parallel in local and in major languages.
J03-3002	2003	99	Using the manually set thresholds for dp and n, we have obtained 100% precision and 68.6% recall in an experiment using STRAND to find English-French Web pages (**Resnik 1999**). **(self citation)**
J03-3002	2003	278	The test set contains 293 of the 326 pairs in **Resniks (1999)** test set. **(self citation)**
J03-3002	2003	275	At different thresholds, Cohens score of agreement (with each of **Resniks (1999)** two judges and their intersection) may be computed for comparison with STRAND, along with recall and precision against a gold standard (for which we use the intersection of the judges: the set of examples B There is some circularity here; the cognates were derived using weighted word pairs from the Bible, then used again in the prior distribution. **(self citation)**
J03-3005	2003	11	Grefenstette and Nioche (2000) and Jones and Ghani (2000) use the Web to generate corpora for languages for which electronic resources are scarce, and **Resnik (1999)** describes a method for mining the Web in order to obtain bilingual texts.
J06-2003	2006	516	Other sources of parallel text, such as parallel translations of the Bible (http://benjamin.umd.edu/parallel/) (**Resnik 1999**) and a collection of Web pages (Resnik, Olsen, and Diab 1999), happened to contain very few occurrences of the near-synonyms of interest.
N07-1045	2007	35	Grefenstette (1999) used the Web for example-based machine translation; Kilgarriff (2001) investigated the type of noise in Web data; Mihalcea and Moldovan (1999) and Agirre and Martinez (2000) used it as an additional resource for word sense disambiguation; **Resnik (1999)** mined the Web for bilingual texts; Turney (2001) used Web frequency counts to compute information retrieval-based mutual-information scores.
P01-1026	2001	7	A sample of these includes methods to extract linguistic resources (Fuji and Ishikawa, 2000; **Resnik, 1999**; Soderland, 1997), retrieve useful information in response to user queries (Etzioni, 1997; McCallum et al. , 1999) and mine/discover knowledge latent in the Web (Inokuchi et al. , 1999).
P03-1058	2003	29	The lack of large-scale parallel corpora no doubt has impeded progress in this direction, although attempts have been made to mine parallel corpora from the Web (**Resnik, 1999**).
P04-1068	2004	73	**Resnik (1999)** addressed the issue of language identification for finding Web pages in the languages of interest.

Fig. 4 Snapshot of the citation summary of Resnik (1999) (Philip Resnik, 1999. "Mining The Web For Bilingual Text," ACL'99.)

61,527 citations are within AAN. This data set has already been used for the development of a reference extraction tool, ParsCit (Councill et al. 2008). They have trained a Conditional Random Field (CRF) to classify each token as "Author" or "Venue" or "Paper Title", etc. in a reference string using manually annotated reference strings as training data.

7.2 Paraphrase acquisition

Previously, we showed in Qazvinian and Radev (2008) that different citations to the same paper they discuss various contributions of the cited paper. Moreover we discussed in Qazvinian and Radev (2011) that the number of factoids (contributions) show asymptotic behavior when the number of citations grow (i.e., the number of contributions of a paper is limited). Therefore, intuitively multiple citations to the same paper may refer to the same contributions of that paper. Since these sentences are written by different authors, they often use different wording to describe the cited factoid. This enables us to use the set of citing sentence pairs that cover the same factoids to create data sets for paraphrase extraction. For example, the sentences below both cite (Turney 2002) and highlight the same aspect of Turney's

Table 11 Sample reference string showing multiple references split over multiple lines

References
David Chiang and Tatjana Scheffler. 2008. Flexible composition and delayed tree-locality. In The Ninth International Workshop on Tree Adjoining Grammars and Related Formalisms (TAG+9)
Aravind K. Joshi and Yves Schabes. 1997. Tree-adjoining grammars. In G. Rozenberg and A. Salo-maa, editors, Handbook of Formal Languages, pages 69â124. Springer.99
Laura Kallmeyer and Maribel Romero. 2004. LTAG semantics with semantic unification. In Proceedings of the 7th International Workshop on Tree-Adjoining Grammars and Related Formalisms (TAG+7), pages 155â162, Vancouver, May
Laura Kallmeyer. 2007. A declarative characterization of different types of multicomponent tree adjoining grammars. In Andreas Witt Georg Rehm and Lothar Lemnitzer, editors, Datenstrukturen fÂur linguistische Ressourcen und ihre Anwendungen, pages 111â120
T. Kasami. 1965. An efficient recognition and syntax algorithm for context-free languages. Technical Report AF-CRL-65-758, Air Force Cambridge Research Laboratory, Bedford, MA

work using slightly different wordings. Therefore, this sentence pair can be considered paraphrases of each other.

> In Turney (2002), an unsupervised learning algorithm was proposed to classify reviews as recommended or not recommended by averaging sentiment annotation of phrases in reviews that contain adjectives or adverbs.
>
> For example, Turney (2002) proposes a method to classify reviews as recommended/not recommended, based on the average semantic orientation of the review.

Similarly, "Eisner (1996) gave a cubic parsing algorithm" and "Eisner (1996) proposed an $O(n^3)$" could be considered paraphrases of each other. Paraphrase annotation of citing sentences consists of manually labeling which sentence consists of what factoids. Then, if two citing sentences consist of the same set of factoids, they are labeled as paraphrases of each other. As a proof of concept, we annotated 25 papers from AAN using the annotation method described above. This data set consisted of 33,683 sentence pairs of which 8,704 are paraphrases (i.e., discuss the same factoids or contributions).

The idea of using citing sentences to create data sets for paraphrase extraction was initially suggested by Nakov et al. (2004) who proposed an algorithm that extracts paraphrases from citing sentences using rules based on automatic named entity annotation and the dependency paths between them.

7.3 Topic modeling

In Hall et al. (2008), this corpus was used to study historical trends in research directions in the field of Computational Linguistics. They also propose a new model to identify which conferences are diverse in terms of topics. They use unsupervised topic modeling using Latent Dirichlet Allocation (Blei et al. 2003) to induce topic clusters. They identify the existence of 46 different topics in AAN and examine the strength of topics over time to identify trends in Computational Linguistics research.

Using the estimated strength of topics over time, they identify which topics have become more prominent and which topics have declined in popularity. They also propose a measure for estimating the diversity in topics at a conference, topic entropy. Using this measure, they identify that EMNLP, ACL, and COLING are increasingly diverse, in that order and are all converging in terms of the topics that they cover.

7.4 Scientific literature summarization

The fact that citing sentences cover different aspects of the cited paper and highlight its most important contributions motivates the idea of using citing sentences to summarize research. The comparison that Elkiss et al. (2008) performed between abstracts and citing sentences suggests that a summary generated from citing sentences will be different and probably more concise and informative than the paper abstract or a summary generated from the full text of the paper. For example, Table 12 shows the abstract of Resnik (1999) and 5 selected sentences that cite it in AAN. We notice that citing sentences contain additional factoids that are not in the abstract, not only ones that summarize the paper contributions, but also those that criticize it (e.g., the last citing sentence in the Table).

Previous work has explored this research direction. Qazvinian and Radev (2008) proposed a method for summarizing scientific articles by building a similarity network of the sentences that cite it, and then applying network analysis techniques to find a set of sentences that covers as much of the paper factoids as possible. Qazvinian et al. (2010) proposed another summarization method that first extracts a number of important keyphrases from the set of citing sentences, and then finds the best subset of sentences that covers as many key phrases as possible.

These works focused on analyzing the citing sentences and selecting a representative subset that covers the different aspects of the summarized article. In recent work, Abu-Jbara and Radev (2011b) raised the issue of coherence and readability in summaries generated from citing sentences. They added pre-processing and post-processing steps to the summarization pipeline. In the pre-processing step, they use a supervised classification approach to rule out irrelevant sentences or fragments of sentences. In the post-processing step, they improve the summary coherence and readability by reordering the sentences, removing extraneous text (e.g. redundant mentions of author names and publication year).

Mohammad et al. (2009) went beyond single paper summarization. They investigated the usefulness of directly summarizing citation texts in the automatic creation of technical surveys. They generated surveys from a set of Question Answering (QA) and Dependency Parsing (DP) papers, their abstracts, and their citation texts. The evaluation of the generated surveys shows that both citation texts and abstracts have unique survey-worthy information. It is worth noting that all the aforementioned research on citation-based summarization used the ACL Anthology Network (AAN) for evaluation.

Table 12 Comparison of the abstract and a selected set of sentences that cite Resnik (1999) work

Abstract	STRAND (Resnik 1998) is a language-independent system for automatic discovery of text in parallel translation on the World WideWeb. This paper extends the preliminary STRAND results by adding automatic language identification, scaling up by orders of magnitude, and formally evaluating performance. The most recent end-product is an automatically acquired parallel corpus comprising 2,491 English-French document pairs, approximately 1.5 million words per language
Selected citing sentences	Many research ideas have exploited the Web in unsupervised or weakly supervised algorithms for natural language processing [e.g., Resnik (1999)]
	Resnik (1999) addressed the issue of language identification for finding Web pages in the languages of interest
	In Resnik (1999), the Web is harvested in search of pages that are available in two languages, with the aim of building parallel corpora for any pair of target languages
	The STRAND system of (Resnik 1999), uses structural markup information from the pages, without looking at their content, to attempt to align them
	Mining the Web for bilingual text (Resnik 1999) is not likely to provide sufficient quantities of high quality data

Table 13 Top authors by research area

Rank	Machine translation	Summarization	Dependency parsing
1	Och, Franz Josef	Lin, Chin-Yew	McDonald, Ryan
2	Koehn, Philipp	Hovy, Eduard H.	Nivre, Joakim
3	Ney, Hermann	McKeown, Kathleen R.	Pereira, Fernando C.N.
4	Della Pietra, Vincent J.	Barzilay, Regina	Nilsson, Jens
5	Della Pietra, Stephen A.	Radev, Dragomir R.	Hall, Johan
6	Brown, Peter F.	Lee, Lillian	Eisner, Jason M.
7	Mercer, Robert L.	Elhadad, Michael	Crammer, Koby
8	Marcu, Daniel	Jing, Hongyan	Riedel, Sebastian
9	Knight, Kevin	Pang, Bo	Ribarov, Kiril
10	Roukos, Salim	Teufel, Simone	Hajič, Jan

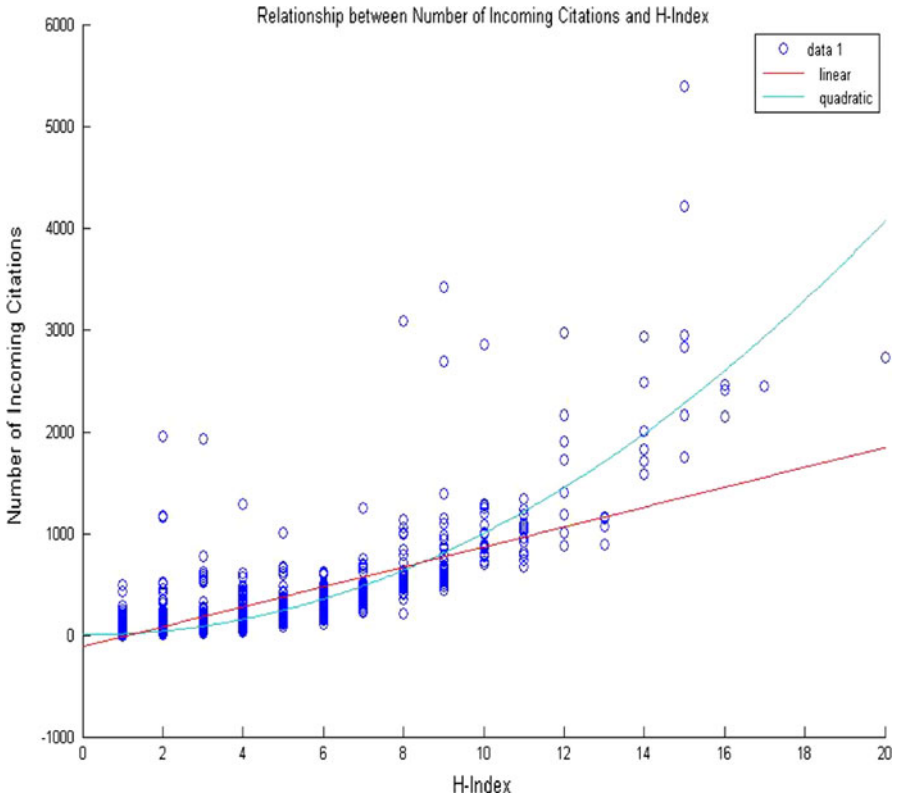

Fig. 5 Relationship between Incoming Citations and h-index

7.5 Finding subject experts

Finding experts in a research area is an important subtask in finding reviewers for publications. We show that using the citation network and the metadata associated with each paper, one can easily find subject experts in any research area.

Table 14 Top 10 outliers for the quadratic function between h-index and incoming citations

Author name	h-index	Incoming citations
Marcinkiewicz, Mary Ann	2	1,950
Zhu, Wei-Jing	2	1,179
Ward, Todd	2	1,157
Santorini, Beatrice	3	1,933
Della Pietra, Vincent J.	9	3,423
Della Pietra, Stephen A.	8	3,080
Brown, Peter F	9	2,684
Dagan, Ido	13	1,155
Moore, Robert C.	13	1,153
Och, Franz Josef	15	5,389

As a proof-of-concept, we performed a simple experiment to find top authors in the following 3 areas "Summarization", "Machine Translation" and "Dependency Parsing". We chose the above three areas because they are some of the most important areas in Natural Language Processing (NLP). We shortlisted papers in each area by searching for papers whose title match the area name. Then we found the top authors by total number of incoming citations to these papers alone. Table 13 lists the top 10 authors in each research area.

7.6 h-index: incoming citations relationship

We performed a simple experiment to find the relationship between the total number of incoming citations and h-index. For the experiment, we chose all the authors who have an *h*-index score of at least 1. We fit a linear function and a quadratic function to the data by minimizing the sum of squared residuals. The fitted curves are shown in Fig. 5. We also measured the goodness of the fit using the sum of the squared residuals. The sum of squared residuals for the quadratic function is equal to 8,240.12 whereas for the linear function it is equal to 10,270.37 which shows that a quadratic function fits the data better as compared to the linear function. Table 14 lists the top 10 outliers for the quadratic function.

7.6.1 Implications of the quadratic relationship

The quadratic relationship between the h-index and total incoming citations adds evidence to the existence of power law in the number of incoming citations (Radev et al. 2009a). It shows that as authors become more successful as shown by higher h-indices they attract more incoming citations. This phenomenon is also known as "the rich get richer" and "preferential attachment" effect.

7.7 Citation context

In Qazvinian and Radev (2010), the corpus is used for extracting context information for citations from scientific articles. Although the citation summaries

have been used successfully for automatically creating summaries of scientific publications in Qazvinian and Radev (2008), additional information consisting of citation context information would be very useful for generating summaries. They report that citation context information in addition to the citation summaries are useful in creating better summaries. They define sentences which contain information about a cited paper but do not explicitly contain the citation as context sentences. For example, consider the following sentence citing (Eisner 1996).

This approach is one of those described in Eisner (1996).

This sentence does not contain any information which can be used for generating summaries. Whereas the surrounding sentences do contain information as follows,

> … In an all pairs approach, every possible pair of two tokens in a sentence is considered and some score is assigned to the possibility of this pair having a (directed) dependency relation. Using that information as building blocks, the parser then searches for the best parse for the sentence. This approach is one of those described in Eisner (1996) …

They model each sentence as a random variable whose value determines its state (context sentence or explicit citation) with respect to the cited paper. They use Markov Random Fields (MRF), a type of graphical model, to perform inference over these random variables. Also, they provide evidence for the usefulness of such citation context information in the generation of surveys of broad research areas.

Incorporating context extraction into survey generation is done in Qazvinian and Radev (2010). They use the MRF technique to extract context information from the datasets used in Mohammad et al. (2009) and show that the surveys generated using the citations as well as context information are better than those generated using abstracts or citations alone. Figure 6 shows a portion of the survey generated from the QA context corpus. This example shows how context sentences add meaningful and survey-worthy information along with citation sentences.

7.8 Temporal analysis of citations

The interest in studying citations stems from the fact that bibliometric measures are commonly used to estimate the impact of a researcher's work (Borgman and Furner 2002; Luukkonen 1992). Several previous studies have performed temporal analysis of citation links (Amblard et al. 2011; Mazloumian et al. 2011; Redner 2005) to see how the impact of research and the relations between research topics evolve overtime. These studies focused on observing how the number of incoming citations to a given article or a set of related articles change over time. However, the number of incoming citations is often not the only factor that changes with time. We believe that analyzing the text of citing sentences allows researchers to observe the change in other dimensions such as the purpose of citation, the polarity of citations, and the research trends. The following subsections discuss some of these dimensions.

Teufel et al. (2006) have shown that the purpose of citation can be determined by analyzing the text of citing sentences. We hypothesize that performing a temporal

... Naturally, our current work on question answering for the reading comprehension task is most related to those of (Hirschman et al. , 1999; Charniak et al. , 2000; Riloffand Thelen, 2000 ; Wang et al. , 2000). **In fact, all of this body of work as well as ours are evaluated on the same set of test stories, and are developed (or trained) on the same development set of stories.** The work of (Hirschman et al. , 1999) initiated this series of work, and it reported an accuracy of 36.3% on answering the questions in the test stories. **Subsequently, the work of (Riloffand Thelen , 2000) and (Chaxniak et al. , 2000) improved the accuracy further to 39.7% and 41%, respectively. However, all of these three systems used handcrafted, deterministic rules and algorithms**...

...**The cross-model comparison showed that the performance ranking of these models was: U-SVM > PatternM > S-SVM > Retrieval-M.** Compared with retrieval-based [Yang et al. 2003], pattern-based [Ravichandran et al. 2002 and Soubbotin et al. 2002], and deep NLP-based [Moldovan et al. 2002, Hovy et al. 2001; and Pasca et al. 2001] answer selection, machine learning techniques are more effective in constructing QA components from scratch. **These techniques suffer, however, from the problem of requiring an adequate number of handtagged question-answer training pairs. It is too expensive and labor intensive to collect such training pairs for supervised machine learning techniques** ...

... **As expected, the definition and person-bio answer types are covered well by these resources.** The web has been employed for pattern acquisition (Ravichandran et al. , 2003), document retrieval (Dumais et al. , 2002), query expansion (Yang et al. , 2003), structured information extraction, and answer validation (Magnini et al. , 2002). **Some of these approaches enhance existing QA systems, while others simplify the question answering task, allowing a less complex approach to find correct answers** ...

Fig. 6 A portion of the QA survey generated by LexRank using the context information

Table 15 Annotation scheme for citation purpose

Comparison	Contrast/comparison in results, method, or goals
Basis	Author uses cited work as basis or starting point
Use	Author uses tools, algorithms, data, or definitions
Description	Neutral description of cited work
Weakness	Limitation or weakness of cited work

analysis of the purpose for citing a paper gives a better picture about its impact. As a proof of concept, we annotated all the citing sentences in AAN that cite the top 10 cited papers from the 1980s with citation purpose labels. The labels we used for annotation are based on Teufel et al.'s annotation scheme and are described in Table 15. We counted the number of times the paper was cited for each purpose in each year since its publication date. Figure 7 shows the change in the ratio of each purpose with time for Shieber's (1985) work on parsing.

The bibliometric measures that are used to estimate the impact of research are often computed based on the number of citations it received. This number is taken as a proxy for the relevance and the quality of the published work. It, however, ignores the fact that citations do not necessarily always represent positive feedback. Many of the citations that a publication receives are neutral citations, and citations that represent negative criticism are not uncommon. To validate this intuition, we annotated about 2,000 citing sentences from AAN for citation polarity. We found that only 30 % of citations are positive, 4.3 % are negative, and the rest are neutral. In another published study, Athar (2011) annotated 8,736 citations from AAN with their polarity and found that only 10 % of citations are positive, 3 % are negative and the rest were all neutral. We believe that considering the polarity of citations when conducting temporal analysis of citations gives more insight about how the way a published work is perceived by the research community over time. As a proof of concept, we annotated the polarity of citing sentences for the top 10 cited papers in AAN that were published in the 1980s. We split the year range of citations into two-year slots and counted the number of positive, negative, and neutral citations

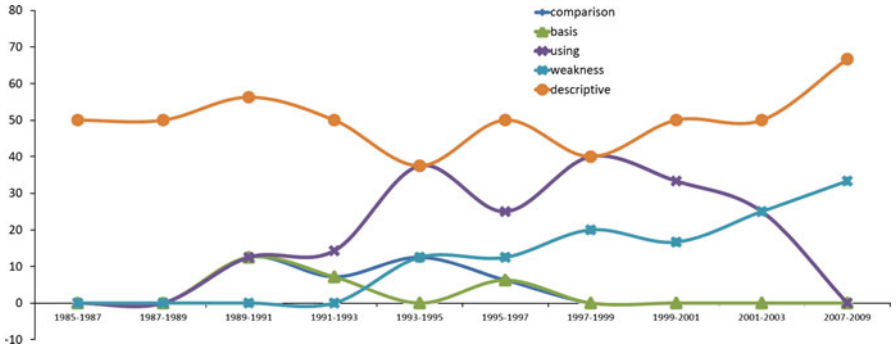

Fig. 7 Change in the citation purpose of Shieber (1985) paper

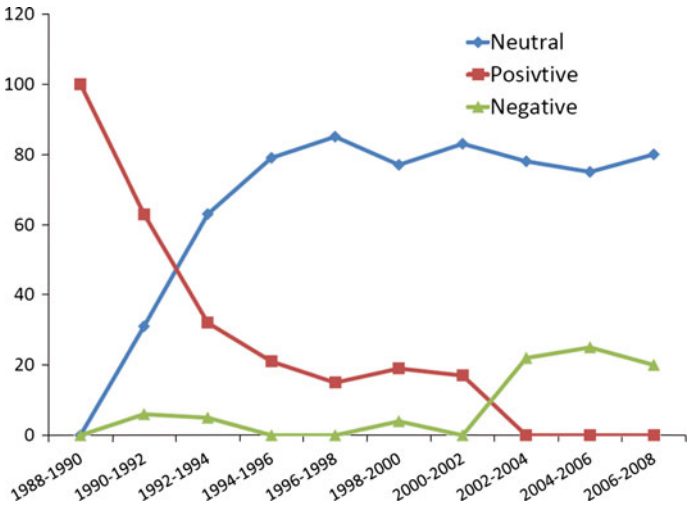

Fig. 8 Change in the polarity of the sentences citing (Church 1988)

that each paper received during that time slot. We observed how the ratios of each category changed overtime. Figure 8 shows the result of this analysis when applied to the work of Church (1988) on part-of-speech tagging.

7.9 Text classification

We chose a subset of papers in 3 topics (Machine Translation, Dependency Parsing, and Summarization) from the ACL anthology. These topics are three main research areas in Natural Language Processing (NLP). Specifically, we collected all papers which were cited by papers whose titles contain any of the following phrases, "Dependency Parsing," "Machine Translation," "Summarization." From this list, we removed all the papers which contained any of the above phrases in their title because this would make the classification task easy. The pruned list contains 1,190

Table 16 A few example papers selected from each research area in the classification data set

ACL-ID	Paper title	Class
W05-0812	Improved HMM Alignment Models for Languages With Scarce Resources	Machine Translation
P07-1111	A Re-Examination of Machine Learning Approaches for Sentence-Level MT Evaluation	Machine Translation
C00-1051	Committee-Based Decision Making in Probabilistic Partial Parsing	Dependency Parsing
C04-1159	Dependency Structure Analysis and Sentence Boundary Detection in Spontaneous Japanese	Dependency Parsing
P88-1020	Planning Coherent Multi-Sentential Text	Summarization

papers. We manually classified each paper into four classes (Dependency Parsing, Machine Translation, Summarization, Other) by considering the full text of the paper. The manually cleaned data set consists of 275 Machine Translation papers, 73 Dependency Parsing papers and 32 Summarization papers for a total of 380 papers. Table 16 lists a few papers from each area.

This data set is slightly different from other text classification data sets in the sense that there are many relational features that are provided for each paper, like textual information, citation information, authorship information, venue information. Recently, There has been a lot of interest in computing better similarity measures for objects by using all the features "together" (Zhou et al. 2008). Since it is very hard to evaluate similarity measures directly, they are evaluated extrinsically using a task for which a good similarity measure directly yields better performance, such as classification.

7.10 Summarizing 30 years of ACL discoveries using citing sentences

The ACL Anthology Corpus contains all the proceedings of the Annual Meeting of the Association of Computational Linguistics (ACL) since 1979. All the ACL papers and their citation links and citing sentences are included in the ACL Anthology Network (ACL). In this section, we show how citing sentences can be used to summarize the most important contributions that have been published in the ACL conference since 1979. We selected the most cited papers in each year and then manually picked a citing sentence that cites a top cited and describes it contribution. It should be noted here that the citation counts we used for ranking papers reflect the number of incoming citations the paper received only from the venues included in AAN. To create the summary, we used citing sentences that cite the same paper at the beginning of the sentence. This is because such citing sentences are often high-quality, concise summaries of the cited work. Table 17 shows the summary of the ACL conference contributions that we created using citing sentences.

Table 17 A citation-based summary of the important contributions published in ACL conference proceedings since 1979

1979	Carbonell (1979) discusses inferring the meaning of new words
1980	Weischedel and Black (1980) discuss techniques for interacting with the linguist/developer to identify insufficiencies in the grammar
1981	Moore (1981) observed that determiners rarely have a direct correlation with the existential and universal quantifiers of first-order logic
1982	Heidorn (1982) provides a good summary of early work in weight-based analysis, as well as a weight-oriented approach to attachment decisions based on syntactic considerations only
1983	Grosz et al. (1983) proposed the centering model which is concerned with the interactions between the local coherence of discourse and the choices of referring expressions
1984	Karttunen (1984) provides examples of feature structures in which a negation operator might be useful
1985	Shieber (1985) proposes a more efficient approach to gaps in the PATR-II formalism, extending Earley's algorithm by using restriction to do top-down filtering
1986	Kameyama (1986) proposed a fourth transition type, Center Establishment (EST), for utterances. e.g., in Bruno was the bully of the neighborhood
1987	Brennan et al. (1987) propose a default ordering on transitions which correlates with discourse coherence
1988	Whittaker and Stenton (1988) proposed rules for tracking initiative based on utterance types; for example, statements, proposals, and questions show initiative, while answers and acknowledgements do not
1989	Church and Hanks (1989) explored tile use of mutual information statistics in ranking co-occurrences within five-word window
1990	Hindle (1990) classified nouns on the basis of co-occurring patterns of subject verb and verb-object pairs
1991	Gale and Church (1991) extract pairs of anchor words, such as numbers, proper nouns (organization, person, title), dates, and monetary information
1992	Pereira and Schabes (1992) establish that evaluation according to the bracketing accuracy and evaluation according to perplexity or cross entropy are very different
1993	Pereira et al. (1993) proposed a soft clustering scheme, in which membership of a word in a class is probabilistic
1994	Hearst (1994) presented two implemented segmentation algorithms based on term repetition, and compared the boundaries produced to the boundaries marked by at least 3 of 7 subjects, using information retrieval metrics
1995	Yarowsky (1995) describes a'semi-unsupervised' approach to the problem of sense disambiguation of words, also using a set of initial seeds, in this case a few high quality sense annotations
1996	Collins (1996) proposed a statistical parser which is based on probabilities of dependencies between head-words in the parse tree

Table 17 continued

1997	Collins (1997)'s parser and its re-implementation and extension by Bikel (2002) have by now been applied to a variety of languages: English (Collins 1999), Czech (Collins et al. 1999), German (Dubey and Keller 2003), Spanish (Cowan and Collins 2005), French (Arun and Keller 2005), Chinese (Bikel 2002) and, according to Dan Bikels web page, Arabic
1998	Lin (1998) proposed a word similarity measure based on the distributional pattern of words which allows to construct a thesaurus using a parsed corpus
1999	Rapp (1999) proposed that in any language there is a correlation between the cooccurrences of words which are translations of each other
2000	Och and Ney (2000) introduce a NULL-alignment capability to HMM alignment models
2001	Yamada and Knight (2001) used a statistical parser trained using a Treebank in the source language to produce parse trees and proposed a tree to string model for alignment
2002	BLEU (Papineni et al. 2002) was devised to provide automatic evaluation of MT output
2003	Och (2003) developed a training procedure that incorporates various MT evaluation criteria in the training procedure of log-linear MT models
2004	Pang and Lee (2004) applied two different classifiers to perform sentiment annotation in two sequential steps: the first classifier separated subjective (sentiment-laden) texts from objective (neutral) ones and then they used the second classifier to classify the subjective texts into positive and negative
2005	Chiang (2005) introduces Hiero, a hierarchical phrase-based model for statistical machine translation
2006	Liu et al. (2006) experimented with tree-to-string translation models that utilize source side parse trees
2007	Goldwater and Griffiths (2007) employ a Bayesian approach to POS tagging and use sparse Dirichlet priors to minimize model size
2008	Huang (2008) improves the re-ranking work of Charniak and Johnson (2005) by re-ranking on packed forest, which could potentially incorporate exponential number of k-best list
2009	Mintz et al. (2009) uses Freebase to provide distant supervision for relation extraction
2010	Chiang (2010) proposes a method for learning to translate with both source and target syntax in the framework of a hierarchical phrase-based system

The top cited paper in each year is found and one citation sentence is manually picked to represent it in the summary

```
id      = {C98-1096}
author  = {Jing, Hongyan; McKeown, Kathleen R.}
title   = {Combining Multiple, Large-Scale Resources in a Reusable Lexicon for Natural
           Language Generation}
Venue   = {International Conference On Computational Linguistics}
year    = {1998}

id      = {J82-3004}
author  = {Church, Kenneth Ward; Patil, Ramesh}
title   = {Coping With Syntactic Ambiguity Or How To Put The Block In  The Box On The
           Table}
venue   = {American Journal Of Computational Linguistics}
year    = {1982}

A00-1001 ==> J82-3002
A00-1002 ==> C90-3057
C08-1001 ==> N06-1007
C08-1001 ==> N06-1008
```

Fig. 9 Sample contents of the downloadable corpus

8 Conclusion

We introduced the ACL Anthology Network (AAN), a manually curated Anthology built on top of the ACL Anthology. AAN, which includes 4 decades of published papers in the field of Computational Linguistics in the ACL community, provides valuable resources for researchers working on various tasks related to scientific data, text, and network mining. These resources include the citation and collaboration networks of more than 18,000 papers from more than 14,000 authors. Moreover AAN includes valuable statistics such as author h-index and PageRank scores. Other manual annotations in AAN include author gender and affiliation annotations, and citation sentence extraction.

In addition to AAN, we also motivated and discussed several different uses of AAN and citing sentences in particular. We showed that citing sentences can be used to analyze the dynamics of research and observe how it trends. We also gave examples on how analyzing the text of citing sentences can give a better understanding of the impact of a researcher's work and how this impact changes over time. In addition, we presented several different applications that can benefit from AAN such as scientific literature summarization, identifying controversial arguments, and identifying relations between techniques, tools and tasks. We also showed how citing sentences from AAN can provide high-quality data for Natural Language Processing tasks such as information extraction, paraphrase extraction, and machine translation. Finally, we used AAN citing sentences to create a citation-based summary of the important contributions included in the ACL conference publication in the past 30 years. The ACL Anthology Network is available to download. The files included in the downloadable package are as follows.

- Text files of the paper: The raw text files of the papers after converting them from pdf to text is available for all papers. The files are named by the corresponding ACL ID.

- Metadata: This file contains all the metadata associated with each paper. The metadata associated with every paper consists of the paper id, title, year, and venue.
- Citations: The paper citation network indicating which paper cites which other paper.
- Database Schema: We have pre-computed the different statistics and stored them in a database which is used for serving the website. The schema of this database is also available for download (Fig. 9).

We also include a large set of scripts which use the paper citation network and the metadata file to output the auxiliary networks and the different statistics.[5] The data set has already been downloaded from 6,930 unique IPs since June 2007. Also, the website has been very popular based on access statistics. There have been nearly 1.1 M hits between April 1, 2009 and March 1, 2010. Most of the hits were searches for papers or authors.

Finally, in addition to AAN, we make Clairlib publicly available to download.[6] The Clairlib library is a suite of open-source Perl modules intended to simplify a number of generic tasks in natural language processing (NLP), information retrieval (IR), and network analysis (NA). Clairlib is in most part developed to work with AAN. Moreover, all of AAN statistics including author and paper network statistics are calculated using the Clairlib library. This library is available for public use for motivated experiments in Sect. 8 as well as to replicate various network statistics in AAN.

As a future direction, we plan to extend AAN to include related conferences and journals including AAAI, SIGIR, ICML, IJCAI, CIKM, JAIR, NLE, JMLR, IR, JASIST, IPM, KDD, CHI, NIPS, WWW, TREC, WSDM, ICSLP, ICASSP, VLDB, and SIGMOD. This corpus, which we refer to as AAN + , includes citations within and between AAN and these conferences. AAN + includes 35,684 papers, with a citation network of 24,006 nodes and 113,492 edges.

References

Abu-Jbara, A., & Radev, D. (2011a). Coherent citation-based summarization of scientific papers. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, Portland, Oregon, USA. Association for Computational Linguistics, pp. 500–509, June.

Abu-Jbara, A., & Radev, D. (2011b). Coherent citation-based summarization of scientific papers. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*. Portland, Oregon, USA: Association for Computational Linguistics, pp. 500–509, June.

Amblard, F., Casteigts, A., Flocchini, P., Quattrociocchi, W., & Santoro, N. (2011). On the temporal analysis of scientific network evolution. In *International conference on computational aspects of social networks (CASoN), 2011*, pp. 169–174, oct.

[5] http://clair.eecs.umich.edu/aan_site2/index.php.

[6] www.clairlib.org/index.php/Download.

Athar, A. (2011). Sentiment analysis of citations using sentence structure-based features. In *Proceedings of the ACL 2011 student session,* pp 81–87, Portland, OR, USA, June. Association for Computational Linguistics.

Bird, S., Dale, R., Dorr, B., Gibson, B., Joseph, M., Kan, M.-Y., Lee, D., et al. (2008). The ACL anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *Language resources and evaluation conference (LREC 08).* Marrakesh, Morocco, May.

Blei, D., Ng, A., & Jordan, M. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Borgman, C. L., & Furner, J. (2002). Scholarly communication and bibliometrics. *Annual Review of Information Science and Technology, 36*(1), 2–72.

Church, K. W. (1988). A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the second conference on applied natural language processing*, pp. 136–143, Austin, Texas, USA, February. Association for Computational Linguistics.

Collins, M. J. (1996). A New Statistical Parser Based On Bigram Lexical Dependencies (ACL, 1996).

Councill, I. G., Lee Giles, C., & Kan, M.-Y. (2008). ParsCit: An open-source CRF reference string parsing package. In *Proceedings of the language resources and evaluation conference (LREC-2008)*, Marrakesh, Morocco.

Eisner, J. (1996). Three new probabilistic models for dependency parsing: An exploration. In *Proceedings of the 34th annual conference of the association for computational linguistics (ACL-96)*, pp. 340–345.

Elkiss, A., Shen, S., Fader, A., Erkan, G., States, D., & Radev, D. (2008). Blind men and elephants: What do citation summaries tell us about a research article? *Journal of the American Society for Information Science Technology, 59*(1), 51–62.

Hall, D., Jurafsky, D., & Manning, C. D. (2008). Studying the History of ideas using topic models. In *EMNLP 2008.*

Luukkonen, T. (1992). Is scientists' publishing behavior rewardseeking? *Scientometrics, 24*, 297–319. doi:10.1007/BF02017913.

Marcus, M. P., Marcinkiewicz, M. A., & Santorini, B. (1993). Building a large annotated corpus of English: The penn treebank (CL, 1993).

Mazloumian, A., Eom, Y.-H., Helbing, D., Lozano, S., & Fortunato, S. (2011). How citation boosts promote scientific paradigm shifts and nobel prizes. *PLoS ONE, 6*(5):e18975, 05.

Mei, Q., & Zhai, C. (2008). Generating impact-based summaries for scientific literature. In *Proceedings of ACL-08: HLT*, pp. 816–824, Columbus, Ohio, June. Association for Computational Linguistics.

Mohammad, S., Dorr, B., Egan, M., Hassan, A., Muthukrishan, P., Qazvinian, V., Radev, D., & Zajic, D. (2009). Using citations to generate surveys of scientific paradigms. In *Proceedings of the North American chapter of the association for computational linguistics—human language technologies (NAACL-HLT-2009)*, May 2009, Boulder, Colorado.

Nakov, P. I., Schwartz, A. S., & Hearst, M. A. (2004). Citances: Citation sentences for semantic analysis of bioscience text. In *Proceedings of the SIGIR04 workshop on search and discovery in bioinformatics*.

Nanba, H., Kando, N., Okumura, M., & Of Information Science. (2000). Classification of research papers using citation links and citation types: Towards automatic review article generation.

Nanba, H., & Okumura, M. (1999). Towards multi-paper summarization using reference information. In *IJCAI'99: Proceedings of the sixteenth international joint conference on artificial intelligence*, pp. 926–931, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Qazvinian, V., & Radev, D. R. (2008). Scientific paper summarization using citation summary networks. In *COLING 2008*, Manchester, UK.

Qazvinian, V., & Radev, D. R. (2010). *Identifying non-explicit citing sentences for citation-based summarization*. ACL.

Qazvinian, V., & Radev, D. R. (2011). Learning from collective human behavior to introduce diversity in lexical choice. In *Proceedings of the 49th Annual Conference of the Association for Computational Linguistics (ACL'11)*, pp. 1098–1108.

Qazvinian, V., Radev, D. R., & Ozgur, A. (2010).*Citation summarization through keyphrase extraction, COLING'10*.

Radev, D. R., Joseph, M., Gibson, B., & Muthukrishnan, P. (2009a). *A bibliometric and network analysis of the field of computational linguistics.* JASIST, 2009.

Radev, D. R., Muthukrishnan, P., & Qazvinian, V. (2009b). 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. Association for Computational Linguistics.

Redner, S. (2005). Citation statistics from 110 years of physical review. *Physics Today, 58*(6), 49–54.

Resnik, P. (1999). Mining the web for bilingual text. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics. Association for Computational Linguistics, (ACL'99)*.

Schäfer, U., Kiefer, B., Spurk, C., Steffen, J., & Wang, R. (2011). The ACL anthology searchbench. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies (ACL HLT 2011), system demonstrations*, pp. 7–13. Portland, OR, USA.

Shieber, S. M. (1985). Using restriction to extend parsing algorithms for complex-feature-based formalisms. In *Proceedings of the 23rd annual meeting of the association for computational linguistics*, pp. 145–152, Chicago, Illinois, USA, July. Association for Computational Linguistics.

Siddharthan, A., & Teufel, S. (2007). Whose idea was this, and why does it matter? Attributing scientific work to citations. In *Proceedings of NAACL/HLT-07*.

Teufel, S. (2007). Argumentative zoning for improved citation indexing. Computing attitude and affect in text. *Theory and Applications,* 159170.

Teufel, S., Siddharthan, A., & Tidhar, D. (2006). Automatic classification of citation function. In *Proceedings of EMNLP-06*.

Turney, P. (2002). Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, (ACL'02)*.

Zhou, D., Zhu, S., Yu, K., Song, X., Tseng, B. L., Zha, H., & Lee Giles, C. (2008). Learning multiple graphs for document recommendations. In *Proceedings of the 17th international world wide web conference (WWW 2008), Beijing, China, 2008.*