The ACL Anthology Network Corpus

Dragomir R. Radev[1,2], Pradeep Muthukrishnan[1], Vahed Qazvinian[1]
[1]Department of Electrical Engineering and Computer Science
[2]School of Information
University of Michigan
{radev,mpradeep,vahed}@umich.edu

Abstract

We introduce the ACL Anthology Network (AAN), a 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.

1 Introduction

The ACL Anthology is one of the most successful initiatives of the ACL. It was initiated by Steven Bird 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. It is available at http://www.aclweb.org/anthology-new/ .

One fundamental problem with the ACL Anthology, however, is the fact that it is just a collection of papers. It doesn't 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 in order to make it possible to compute such statistics.

In addition, we were able to use the annotated data for extracting citation summaries of all papers in the collection and we also annotated each paper by the gender of the authors (and are currently in the process of doing similarly for their institutions) in the goal of creating multiple gold standard data sets for training automated systems for performing such tasks.

2 Curation

The ACL Anthology includes 13,739 papers (excluding book reviews and posters). Each of the papers was converted from pdf to text using an OCR tool (www.pdfbox.org). After this conversion, we extracted the references semi-automatically using string matching. The above process outputs all the references as a single block so we then 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 Figure 1. The matched references were stored together to produce the citation network. References to publications outside of the AAN were recorded but not included in the network.

In order to fix the issue of wrong author names and multiple author identities we had to perform a lot of 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". There were a few spelling mistakes, like "Madeleine Bates" was misspelled as "Medeleine Bates".

Finally, many papers included incorrect titles in their citation sections. Some used the wrong years and/or venues as well.

Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries, ACL-IJCNLP 2009, pages 54–61,
Suntec, Singapore, 7 August 2009.

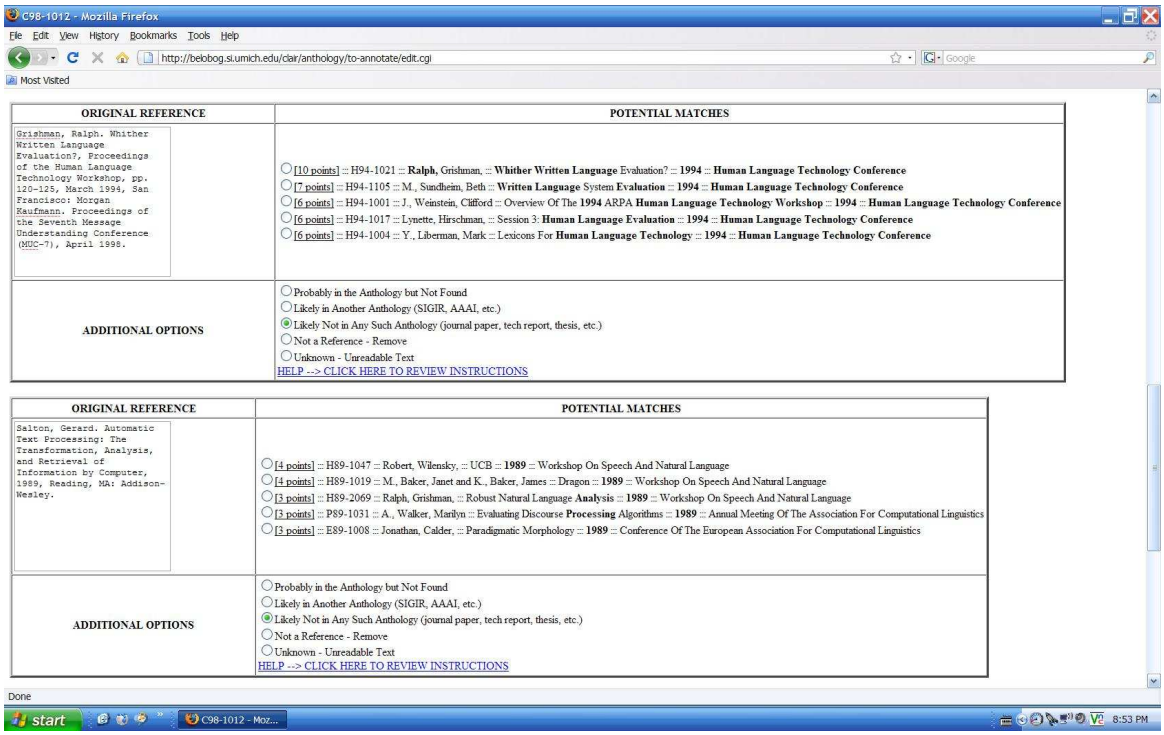


Figure 1: CGI interface used for matching new references to existing papers

Figure 2: Snapshot of the different statistics computed for an author

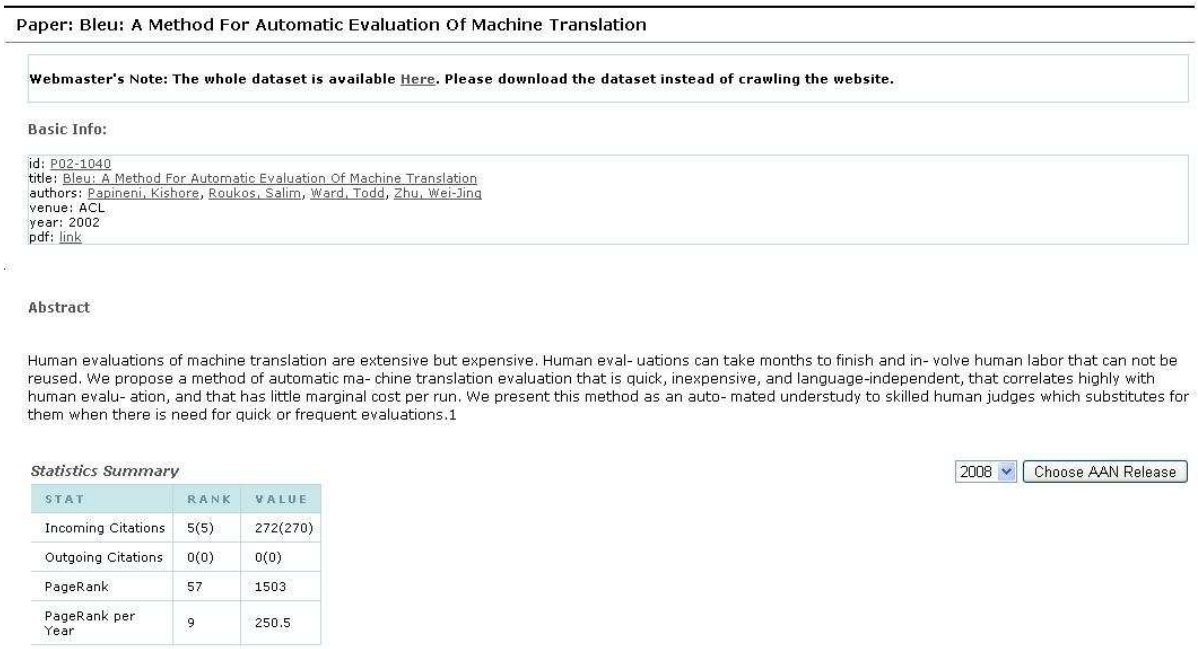


Figure 3: Snapshot of the different statistics for a paper

3 Statistics

Using the metadata and the citations extracted after curation, we have built three different networks.

The paper citation network is a directed network with each node representing a paper labeled with an ACL ID number and the edges representing a citation within that paper to another paper represented by an ACL ID. The paper citation network consists of 13,739 papers and 54,538 citations.

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 11,180 unique authors and 332,815 edges (196,905 edges if duplicates are removed).

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

Table 1 shows some brief statistics about the first two releases of the data set (2006 and 2007). Table 2 describes the most current release of the data set (from 2008).

2006			
	Paper citation network	Author citation network	Author collaboration network
n	8898	7849	7849
m	8765	137,007	41,362
2007			
	Paper citation network	Author citation network	Author collaboration network
n	9767	9421	9421
m	44,142	158,479	45,878

Table 1: Growth of citation volume

	Paper Citation Network	**Author Citation Network**	**Author Collaboration Network**
Nodes	13,739	10,409	10,409
Edges	54,538	195,505	57,614
Diameter	22	10	20
Average	9.34	43.11	11.07

Degree			
Largest Connected Component	11,409	9061	7910
Watts Strogatz clustering coefficient	0.18	0.46	0.65
Newman clustering coefficient	0.07	0.14	0.36
clairlib avg. directed shortest path	5.91	3.32	5.87
Ferrer avg. directed shortest path	5.35	3.29	4.66
harmonic mean geodesic distance	63.93	5.47	9.40
harmonic mean geodesic distance with self-loops counted	63.94	5.47	9.40

Table 2: Network Statistics of the citation and collaboration network. The remaining authors (11,180-10,409) are not cited and are therefore removed from the network analysis

	Paper Citation Network	**Author Citation Network**	**Author Collaboration Network**
In-degree Stats			
Power Law Exponent	2.50	2.20	3.17
Power Law Relationship?	No	No	No
Newman Power Law exponent	2.00	1.55	2.18
Out-degree stats			
Power Law Exponent	3.70	2.56	3.17
Power Law Relationship?	No	No	No
Newman Power Law exponent	2.12	1.54	2.18
Total Degree Stats			
Power Law Exponent	2.72	2.27	3.17
Power Law Relationship?	No	No	No
Newman Power Law exponent	1.81	1.46	2.18

Table 3: Degree Statistics of the citation and collaboration networks

A lot of different statistics have been computed based on the data set release in 2007 by Radev et al. The statistics include PageRank scores which eliminate PageRank's inherent bias towards older papers, Impact factor, correlations between different measures of impact like H-Index, total number of incoming citations, PageRank. They 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 Sample rankings

This section shows some of the rankings that were computed using AAN.

Rank	*Icit*	*Title*
1	590	Building A Large Annotated Corpus Of English: The Penn Treebank
2	444	The Mathematics Of Statistical Machine Translation: Parameter Estimation
3	324	Attention Intentions And The Structure Of Discourse
4	271	A Maximum Entropy Approach To Natural Language Processing
5	270	Bleu: A Method For Automatic Evaluation Of
6	246	A Maximum-Entropy-Inspired Parser
7	230	A Stochastic Parts Program And Noun Phrase Parser For Unrestricted Text
8	221	A Systematic Comparison Of Various Statistical Alignment
9	211	A Maximum Entropy Model For Part-Of-Speech Tagging
10	211	Three Generative Lexicalized Models For Statistical Parsing

Table 4: Papers with the most incoming citations (icit)

Rank	*PR*	*Title*
1	1099.1	A Stochastic Parts Program And Noun Phrase Parser For Unrestricted Text
2	943.8	Finding Clauses In Unrestricted Text By Finitary And Stochastic Methods
3	568.8	A Stochastic Approach To
4	543.1	A Statistical Approach To Machine Translation
5	414.1	Building A Large Annotated Corpus Of English: The Penn Treebank
6	364.9	The Mathematics Of Statistical Machine Translation: Parameter Estimation
7	362.2	The Contribution Of Parsing To Prosodic Phrasing In An Experimental Text-To-Speech System
8	301.6	Attention Intentions And The Structure Of Discourse
9	250.5	Bleu: A Method For Automatic Evaluation Of Machine Translation
10	242.5	A Maximum Entropy Approach To Natural Language

Table 5: Papers with highest PageRank (PR) scores

It must be noted that the PageRank scores are not accurate because of the lack of citations outside AAN. Specifically, out of the 155,858 total number of citations, only 54,538 are within AAN.

Rank	*Icit*	*Author Name*
1 (1)	3886 (3815)	Och, Franz Josef
2 (2)	3297 (3119)	Ney, Hermann
3 (3)	3067 (3049)	Della Pietra, Vincent J.
4 (5)	2746 (2720)	Mercer, Robert L.
5 (4)	2741 (2724)	Della Pietra, Stephen
6 (6)	2605 (2589)	Marcus, Mitchell P.
7 (8)	2454 (2407)	Collins, Michael John
8 (7)	2451 (2433)	Brown, Peter F.
9 (9)	2428 (2390)	Church, Kenneth Ward
10 (10)	2047 (1991)	Marcu, Daniel

Table 6: Authors with most incoming citations (the values in parentheses are using non-self- citations)

Rank	*h*	*Author Name*
1	18	Knight, Kevin
2	16	Church, Kenneth Ward
3	15	Manning, Christopher D.
3	15	Grishman, Ralph
3	15	Pereira, Fernando C. N.
6	14	Marcu, Daniel
6	14	Och, Franz Josef
6	14	Ney, Hermann
6	14	Joshi, Aravind K.
6	14	Collins, Michael John

Table 7: Authors with the highest h-index

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

Table 8: Authors with the least average shortest path (ASP) length in the author collaboration network

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 (see Figure 4).

Closest Words/Phrase

	WORD	TF-IDF
1	alignment	3060.28788645363
2	translation	1609.64150036477
3	bleu	1270.66151594014
4	rouge	1131.61343683879
5	och	1070.2577306796
6	ney	1032.93379864255
7	alignments	938.646118573016
8	translations	779.35942419005
9	prime	606.568302266622
10	training	562.098194260184

Figure 4: Snapshot of the related phrases for Franz Josef Och

6 Citation summaries

The citation summary of an article, *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, 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 following example shows the citation summary extracted for "Koo, Terry, Carreras, Xavier, Collins, Michael John, Simple Semi-supervised Dependency Parsing". The citation summary of (Koo et al., 2008) mentions KCC08, dependency parsing, and the use of word clustering in semi-supervised NLP.

C08-1051 1 7:191 Furthermore, recent studies revealed that word clustering is useful for semi-supervised learning in NLP (Miller et al., 2004; Li and McCallum, 2005; Kazama and Torisawa, 2008; Koo et al., 2008).

D08-1042 2 78:214 There has been a lot of progress in learning dependency tree parsers (McDonald et al., 2005; Koo et al., 2008; Wang et al., 2008).

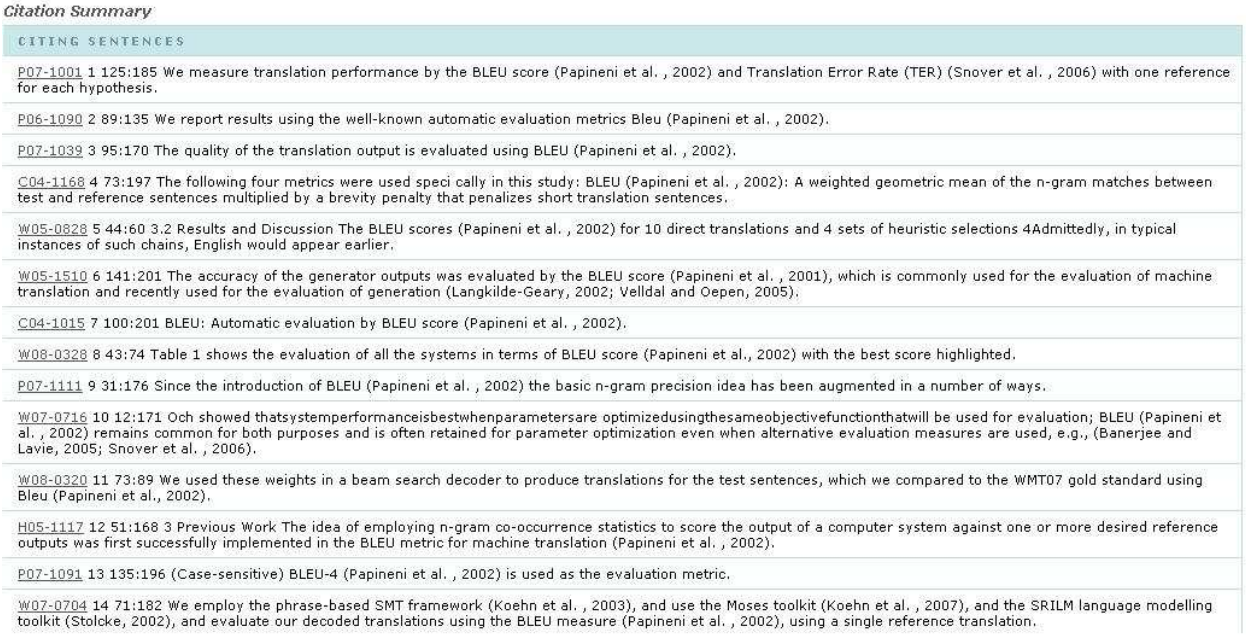
W08-2102 3 194:209 The method shows improvements over the method described in (Koo et al., 2008), which is a state-of-the-art second-order dependency parser similar to that of (McDonald and Pereira, 2006), suggesting that the incorporation of constituent structure can improve dependency accuracy.

W08-2102 4 32:209 The model also recovers dependencies with significantly higher accuracy than state-of-the-art dependency parsers such as (Koo et al., 2008; McDonald and Pereira, 2006).

W08-2102 5 163:209 KCC08 unlabeled is from (Koo et al., 2008), a model that has previously been shown to have higher accuracy than (McDonald and Pereira, 2006).

W08-2102 6 164:209 KCC08 labeled is the labeled dependency parser from (Koo et al., 2008); here we only evaluate the unlabeled accuracy.

Figure 5: Sample citation summary



Citation Summary

CITING SENTENCES

P07-1001 1 125:185 We measure translation performance by the BLEU score (Papineni et al. , 2002) and Translation Error Rate (TER) (Snover et al. , 2006) with one reference for each hypothesis.

P06-1090 2 89:135 We report results using the well-known automatic evaluation metrics Bleu (Papineni et al. , 2002).

P07-1039 3 95:170 The quality of the translation output is evaluated using BLEU (Papineni et al. , 2002).

C04-1168 4 73:197 The following four metrics were used speci cally in this study: BLEU (Papineni et al. , 2002): A weighted geometric mean of the n-gram matches between test and reference sentences multiplied by a brevity penalty that penalizes short translation sentences.

W05-0828 5 44:60 3.2 Results and Discussion The BLEU scores (Papineni et al. , 2002) for 10 direct translations and 4 sets of heuristic selections 4Admittedly, in typical instances of such chains, English would appear earlier.

W05-1510 6 141:201 The accuracy of the generator outputs was evaluated by the BLEU score (Papineni et al. , 2001), which is commonly used for the evaluation of machine translation and recently used for the evaluation of generation (Langkilde-Geary, 2002; Velldal and Oepen, 2005).

C04-1015 7 100:201 BLEU: Automatic evaluation by BLEU score (Papineni et al. , 2002).

W08-0328 8 43:74 Table 1 shows the evaluation of all the systems in terms of BLEU score (Papineni et al., 2002) with the best score highlighted.

P07-1111 9 31:176 Since the introduction of BLEU (Papineni et al. , 2002) the basic n-gram precision idea has been augmented in a number of ways.

W07-0716 10 12:171 Och showed thatsystemperformanceisbestwhenparametersare optimizedusingthesameobjectivefunctionthatwill be used for evaluation; BLEU (Papineni et al. , 2002) remains common for both purposes and is often retained for parameter optimization even when alternative evaluation measures are used, e.g., (Banerjee and Lavie, 2005; Snover et al. , 2006).

W08-0320 11 73:89 We used these weights in a beam search decoder to produce translations for the test sentences, which we compared to the WMT07 gold standard using Bleu (Papineni et al., 2002).

H05-1117 12 51:168 3 Previous Work The idea of employing n-gram co-occurrence statistics to score the output of a computer system against one or more desired reference outputs was first successfully implemented in the BLEU metric for machine translation (Papineni et al. , 2002).

P07-1091 13 135:196 (Case-sensitive) BLEU-4 (Papineni et al. , 2002) is used as the evaluation metric.

W07-0704 14 71:182 We employ the phrase-based SMT framework (Koehn et al. , 2003), and use the Moses toolkit (Koehn et al. , 2007), and the SRILM language modelling toolkit (Stolcke, 2002), and evaluate our decoded translations using the BLEU measure (Papineni et al. , 2002), using a single reference translation.

Figure 6: Snapshot of the citation summary for a paper

The citation text that we have extracted for each paper is a good resource to generate summaries of the contributions of that paper. We have previously developed systems using clustering the similarity networks to generate short, and yet informative, summaries of individual papers (Qazvinian and Radev 2008), and more general scientific topics, such as Dependency Parsing, and Machine Translation (Radev et al. 2009) .

7 Gender annotation

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.

8 Downloads

The following files can be downloaded:

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, venue.

Citations: The paper citation network indicating which paper cites which other paper.

Figure 7 includes some examples.

```
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

Figure 7: Sample contents of the downloadable corpus

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.

The scripts are documented here: http://clair.si.umich.edu/ .The data set has already been downloaded from 2,775 unique IPs since June 2007. Also, the website has been very popular based on access statistics. There have been more than 2M accesses in 2009.

References

Vahed Qazvinian and Dragomir R. Radev. Scientific paper summarization using citation summary networks. In COLING 2008, Manchester, UK, 2008.

Dragomir R. Radev, Mark Joseph, Bryan Gibson, and Pradeep Muthukrishnan. A Bibliometric and Network Analysis of the Field of Computational Linguistics. JASIST, 2009 to appear.