Examining Citations of Natural Language Processing Literature

Saif M. Mohammad
National Research Council Canada
Ottawa, Canada
saif.mohammad@nrc-cnrc.gc.ca.

Abstract

We extracted information from the ACL Anthology (AA) and Google Scholar (GS) to examine trends in citations of NLP papers. We explore questions such as: how well cited are papers of different types (journal articles, conference papers, demo papers, etc.)? how well cited are papers from different areas of within NLP? etc. Notably, we show that only about 56% of the papers in AA are cited ten or more times. CL Journal has the most cited papers, but its citation dominance has lessened in recent years. On average, long papers get almost three times as many citations as short papers; and papers on *sentiment classification*, *anaphora resolution*, and *entity recognition* have the highest median citations. The analyses presented here, and the associated dataset of NLP papers mapped to citations, have a number of uses including: understanding how the field is growing and quantifying the impact of different types of papers.

1 Introduction

The origins of Natural Language Processing (NLP) go back to the earliest work in Computer Science—when Alan Turing published his seminal paper exploring whether machines can think, and proposed what is now known as the *Turing test* (Turing, 1950, 2009). A crucial factor in the evolution of NLP as a field of study in its own right was the formation of the Association for Computational Linguistics (ACL) in 1962, and the first ACL conference in 1965.[1] Today NLP is a broad interdisciplinary field with a growing number of researchers from Computer Science, Linguistics, Information Science, Psychology, Social Sciences, Humanities, and more joining its ranks.

[1]One can make a distinction between NLP and Computational Linguistics; however, for this work, we will consider them to be synonymous. Also, ACL was originally named the Association for Machine Translation and Computational Linguistics (AMTCL). It was changed to ACL in 1968.

Organizations such as ACL, ELRA, and AFNLP publish peer-reviewed NLP papers that include both journal articles and conference proceedings. Historically, the need for a faster review process has made conference proceedings the dominant form of published research in Computer Science and NLP. With time, the conferences and the types of papers they publish, have evolved. Some conferences, such as EMNLP and ACL, are highly competitive, while others, such as most workshops and LREC, deliberately choose to keep more generous acceptance rates. The publications themselves can be of different types: journal articles, conference papers, short papers, system demonstration papers, shared task papers, workshop papers, etc. New ideas and paradigms have evolved: for example, the rise of statistical NLP in the 1990s and deep learning in the 2010s. With the dawn of a new decade and NLP research becoming more diverse and more popular than it ever has been, this work looks back at the papers already published to identify broad trends in their impact on subsequent scholarly work.

Commonly used metrics of research impact on subsequent scholarly work are derived from citations including: number of citations, average citations, h-index, relative citation ratio, and impact factor (Bornmann and Daniel, 2009). However, the number of citations is not always a reflection of the quality or importance of a piece of work. Note also that there are systematic biases that prevent certain kinds of papers from accruing citations, especially when the contributions of a piece of work are atypical or in an area where the number of scientific publications is low. Furthermore, the citation process can be abused, for example, by egregious self-citations (Ioannidis et al., 2019). Nonetheless, given the immense volume of scientific literature, the relative ease with which one can track citations using services such as Google Scholar (GS), and given the lack of other easily applicable and effec-

tive metrics, citation analysis is an imperfect but useful window into research impact.

Thus citation metrics are often a factor when making decisions about funding research and hiring scientists. Citation analysis can also be used to gauge the influence of outside fields on one's field and the influence of one's field on other fields. Therefore, it can be used to determine the relationship of a field with the wider academic community.

As part of a broader project on analyzing NLP Literature, we extracted and aligned information from the ACL Anthology (AA) and Google Scholar to create a dataset of tens of thousands of NLP papers and their citations (Mohammad, 2020b, 2019).[2] In this paper, we describe work on examining the papers and their citations to identify broad trends within NLP research—overall, across paper types, across publication venues, over time, and across research areas within NLP. Notably, we explored questions such as: how well cited are papers of different types (journal articles, conference papers, demo papers, etc.)? how well cited are papers published in different time spans? how well cited are papers from different areas of research within NLP? etc. The dataset and the analyses have many uses including: understanding how the field is growing; quantifying the impact of different types of papers on subsequent publications; and understanding the impact of various conferences and journals. Perhaps most importantly, though, they serve as a record of the state of NLP literature in terms of citations. All of the data and interactive visualizations associated with this work are freely available through the project homepage.[3]

2 Background and Related Work

The ACL Anthology is a digital repository of public domain, free to access, articles on NLP.[4] It includes papers published in the family of ACL conferences as well as in other NLP conferences such as LREC and RANLP.[5] As of June 2019, it provided access to the full text and metadata for ∼50K articles published since 1965 (the year of the first ACL conference). It is the largest single source of scientific literature on NLP. Various subsets of AA have been used in the past for a number of tasks including: the study of citation patterns and intent (Pham and Hoffmann, 2003; Aya et al., 2005; Teufel et al., 2006; Mohammad et al., 2009; Nanba et al., 2011; Zhu et al., 2015; Radev et al., 2016), generating summaries of scientific articles (Qazvinian et al., 2013), and creating corpora of scientific articles (Bird et al., 2008; Mariani et al., 2018). Perhaps the work closest to ours is that by Anderson et al. (2012), who examine papers from 1980 to 2008 to track the ebb and flow of topics within NLP, the influence of subfields on each other, and the influence of researchers from outside NLP. However, that work did not examine trends in the citations of NLP papers.

Google Scholar is a free web search engine for academic literature.[6] Through it, users can access the metadata associated with an article such as the number of citations it has received. Google Scholar does not provide information on how many articles are included in its database. However, scientometric researchers estimated that it included about 389 million documents in January 2018 (Gusenbauer, 2019)—making it the world's largest source of academic information. Thus, there is growing interest in the use of Google Scholar information to draw inferences about scholarly research in general (Howland et al., 2009; Orduña-Malea et al., 2014; Khabsa and Giles, 2014; Mingers and Leydesdorff, 2015; Martín-Martín et al., 2018) and on scholarly impact in particular (Priem and Hemminger, 2010; Yogatama et al., 2011; Bulaitis, 2017; Ravenscroft et al., 2017; Bos and Nitza, 2019; Ioannidis et al., 2019). This work examines patterns of citations of tens of thousands of NLP papers, both overall and across paper types, venues, and areas of research.

3 Data

We now briefly describe how we extracted information from the ACL Anthology and Google Scholar to facilitate the citation analysis. (Further details about the dataset, as well as an analysis of the volume of research in NLP over the years, are available in Mohammad (2020b).) We aligned the information across AA and GS using the paper title, year of publication, and first author last name.

[2] In separate work we have used the NLP Scholar data to explore gender gaps in Natural Language Processing research; especially, disparities in authorship and citations (Mohammad, 2020a). We have also developed an interactive visualization tool that allows users to search for relevant related work in the ACL Anthology Mohammad (2020c).

[3] http://saifmohammad.com/WebPages/nlpscholar.html

[4] https://www.aclweb.org/anthology/

[5] ACL licenses its papers with a Creative Commons Attribution 4.0 International License.

[6] https://scholar.google.com

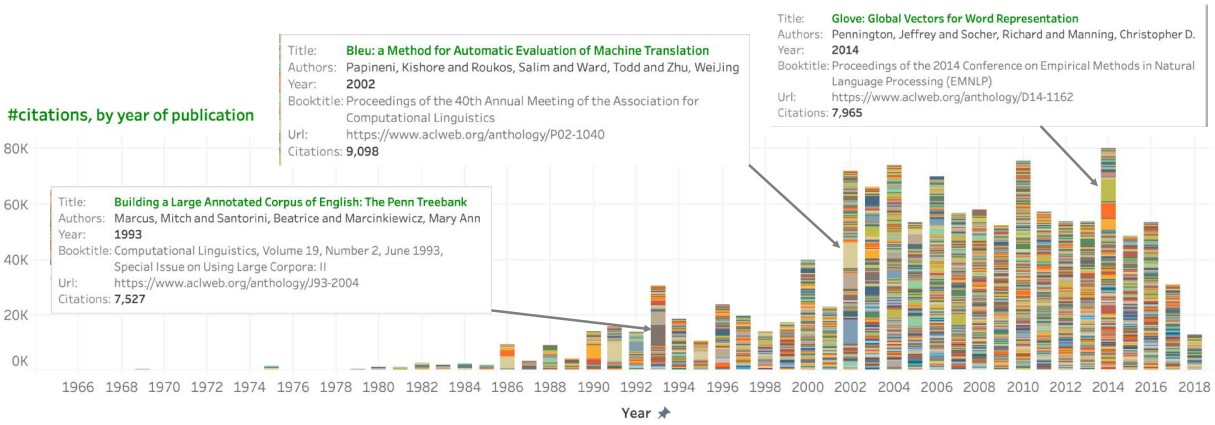

Figure 1: A timeline graph of citations received by papers published in each year. Colored segments correspond to papers; the height of a segment is proportional to the number of citations. Hovering over a paper shows metadata.

3.1 ACL Anthology Data

The ACL Anthology provides access to its data through its website and a github repository (Gildea et al., 2018).[7] We extracted paper title, names of authors, year of publication, and venue of publication from the repository.[8]

As of June 2019, AA had ∼50K entries; however, this includes forewords, schedules, etc. that are not truly research publications. After discarding them we are left with a set of 44,894 papers.[9]

3.2 Google Scholar Data

Google Scholar does not provide an API to extract information about the papers. This is likely because of its agreement with publishing companies that have scientific literature behind paywalls (Martín-Martín et al., 2018). We extracted citation information from Google Scholar profiles of authors who published at least three papers in the ACL Anthology. A Google Scholar Profile page is a user-created page where authors can include their papers (along with the GS-provided citation information for the papers). Scraping author profile pages is explicitly allowed by GS's robots exclusion standard. This is also how past work has studied Google Scholar (Khabsa and Giles, 2014; Orduña-Malea et al., 2014; Martín-Martín et al., 2018).

We collected citation information for 1.1 million papers in total. We will refer to this dataset as *GScholar-NLP*. Note that GScholar-NLP includes citation counts not just for NLP papers, but also for non-NLP papers published by the authors. GScholar-NLP includes 32,985 of the 44,894 papers in AA (about 74%). We will refer to this subset of the ACL Anthology papers as AA′. The citation analyses presented in this paper are on AA′. Future work will analyze both AA′ and GScholar-NLP to determine influences of other fields on NLP.

4 Examining Citations of NLP Papers

We use data extracted from the ACL Anthology and Google Scholar to examine trends in citations through a series of questions.

Q1. How many citations have the AA′ papers received? How is that distributed among the papers published in various years?

A. ∼1.2 million citations (as of June 2019). Figure 1 shows the screenshot of an interactive timeline graph where each year has a bar with height corresponding to the number of citations received by papers published in that year. Further, the bar has colored segments corresponding to each of the papers; the height of a segment is proportional to the number of citations the paper has received. Thus it is easy to spot the papers that received a large number of citations. Hovering over individual papers reveals additional metadata.

[7] https://www.aclweb.org/anthology/ https://github.com/acl-org/acl-anthology

[8] Multiple authors can have the same name and the same authors may use multiple variants of their names in papers. The AA volunteer team handles such ambiguities using both semi-automatic and manual approaches (fixing some instances on a case-by-case basis). Additionally, the AA repository includes a file that has canonical forms of author names.

[9] We used simple keyword searches for terms such as *foreword, invited talk, program, appendix* and *session* in the title to pull out entries that were likely to not be research publications. These were then manually examined to verify that they did not contain any false positives.

Discussion: With time, not only have the number of papers grown, but also the number of high-citation papers. We see a marked jump in the 1990s over the previous decades, but the 2000s are the most notable in terms of the high number of citations. The 2010s papers will likely surpass the 2000s papers in the years to come.

Q2. How well cited are individual AA′ papers, as in, what is the average number of citations, what is the median, what is the distributison of citations? How well cited are the different types of papers: journal papers, main conference papers, workshop papers, etc.?

A. In this and all further analyses, we do not include AA′ papers published in 2017 or later (to allow for at least 2.5 years for the papers to collect citations). There are 26,949 AA′ papers that were published from 1965 to 2016. Figure 2 shows box and whisker plots for: all of these papers (on the left) and for individual paper types (on the right). The whiskers are at a distance of 1.5 times the inter-quartile length. The average number of citations are indicated with the horizontal green dotted lines. Creating a separate class for "Top-tier Conference" is somewhat arbitrary, but it helps make certain comparisons more meaningful. For this work, we consider ACL, EMNLP, NAACL, COLING, and EACL as top-tier conferences based on low acceptance rates and high citation metrics, but certainly other groupings are also reasonable.

Discussion: Overall, the median citation count is 12. 75% of the papers have 34 or fewer citations. The average number of citations (45) is markedly higher than the median (12); this is because of a small number highly cited papers.

When comparing different types of papers, we notice a large difference between journal papers and the rest. Even though the number of journal papers in AA (and AA′) is very small (about 2.5%), these papers have the highest median and average citations (55 and 204, respectively). Top-tier conferences come next, followed by other conferences. The differences between each of these pairs is statistically significant (Kolmogorov–Smirnov (KS) test, $p < .01$).[10] Interestingly, the workshop papers and the shared task papers have higher medians

[10]KS is a non-parametric test that can be applied to compare distributions without needing to make assumptions about the nature of the distributions. Since the citations data is not normally distributed, KS is especially well suited.

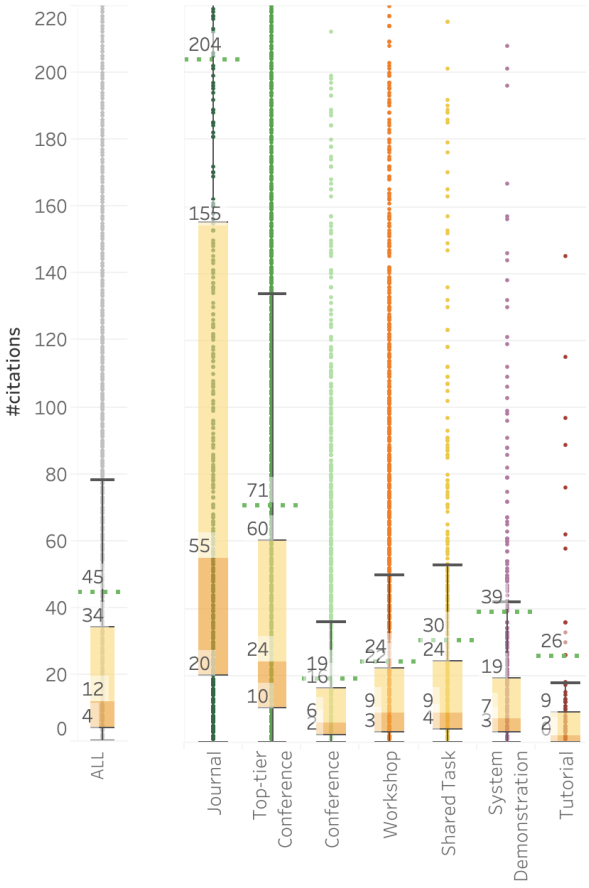

Figure 2: Citation box plots for papers published 1965–2016: overall and by type.

and averages than the non-top-tier conferences. These differences are also significant (KS, $p < .01$).

Q3. How well cited are recent AA′ papers: say those published in the last decade (2010–2016)? How well cited are papers that were all published in the same year, say 2014? Are the citation distributions for individual years very different from those for larger time spans, say 2010–2016? Also, how well cited are papers 5 years after they are published?

A. The top of Figure 3 shows citation box plots for 2010–2016; the bottom shows plots for papers published in 2014.

Discussion: Observe that, in general, these numbers are markedly lower than the those in Figure 2. That is expected as these papers have had less time to accrue citations.

Observe that journal papers again have the highest median and average; however, the gap between journals and top-tier conferences has reduced considerably. The shared task papers have a signifi-

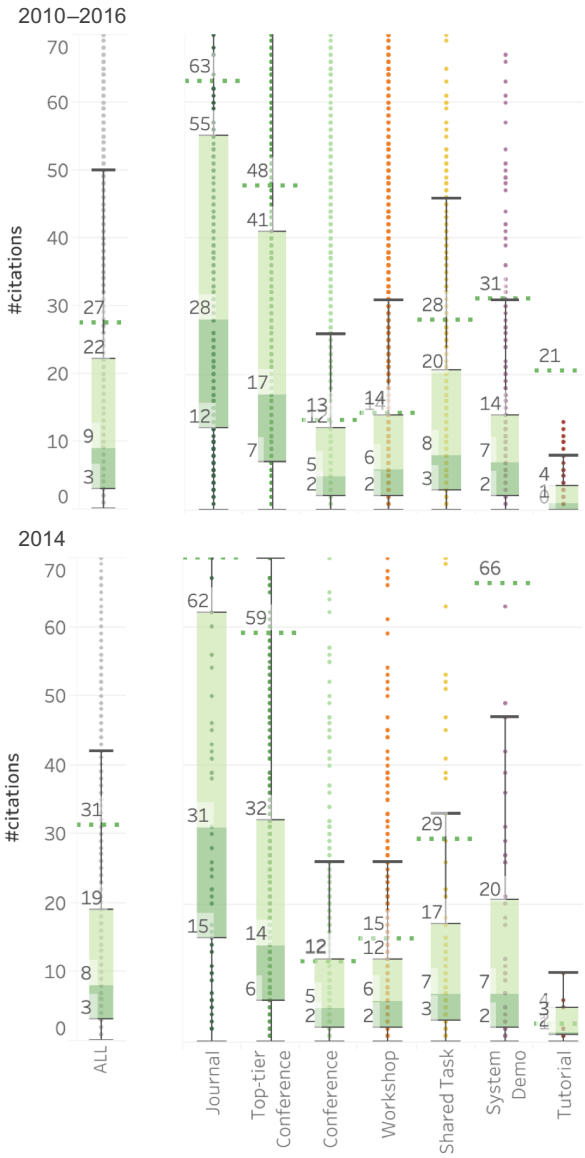

Figure 3: Citation box plots for papers: published 2010–2016 (top) and published in 2014 (bottom).

cantly higher average than workshop and non-top-tier conferences. Examining the data revealed that many of the task description papers and the competition winning systems' system-description papers received a large number of citations (while the majority of the other system description papers received much lower citations). Shared tasks have also been particularly popular in the 2010s compared to earlier years.

The plots for 2014 (bottom of Figure 3) are similar to that of 2010–2016. (Although, system demo papers published in that year are better cited

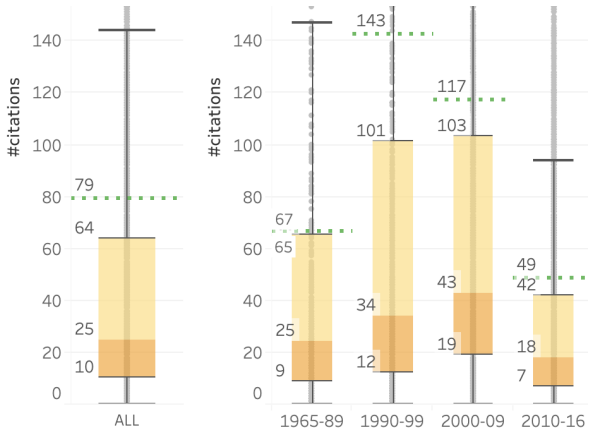

Figure 4: Citation box plots for journal articles and top-tier conference papers from various time spans.

than the larger set from the 2010–2016 period.) This plot also gives an idea of citation patterns for papers 5 years after they have been published.

Q4. If we only consider journal papers and top-tier conferences, how well cited are papers from various time spans?

A. Figure 4 shows the numbers for four time spans.

Discussion: Observe that the 1990s and the 2000s have markedly higher medians and averages than other time periods. The early 1990s, which have the highest average, were an interesting period for NLP with the emergence of statistical approaches (especially from speech processing) and the use of data from the World Wide Web. The 2000–2010 period, which saw an intensification of the statistical data-driven approaches, is notable for the highest median. The high average in the 1990s is likely because of some seminal papers that obtained a very high number of citations. (Also the 1990's had fewer papers than the 2010s, and thus the average is impacted more by the very high-citation papers.) The drop off in the average and median for recent papers is largely because they have not had as much time to collect citations.

Q5. *How well cited are papers from individual NLP venues?*

A. Figure 5 (top) shows the citation box plots for 1965–2016 papers from individual venues. The plots for workshops, system, demos, shared tasks, and tutorials are shown as well for ease of comparison. Figure 5 (bottom) shows the same box plots for 2010–2016 papers.

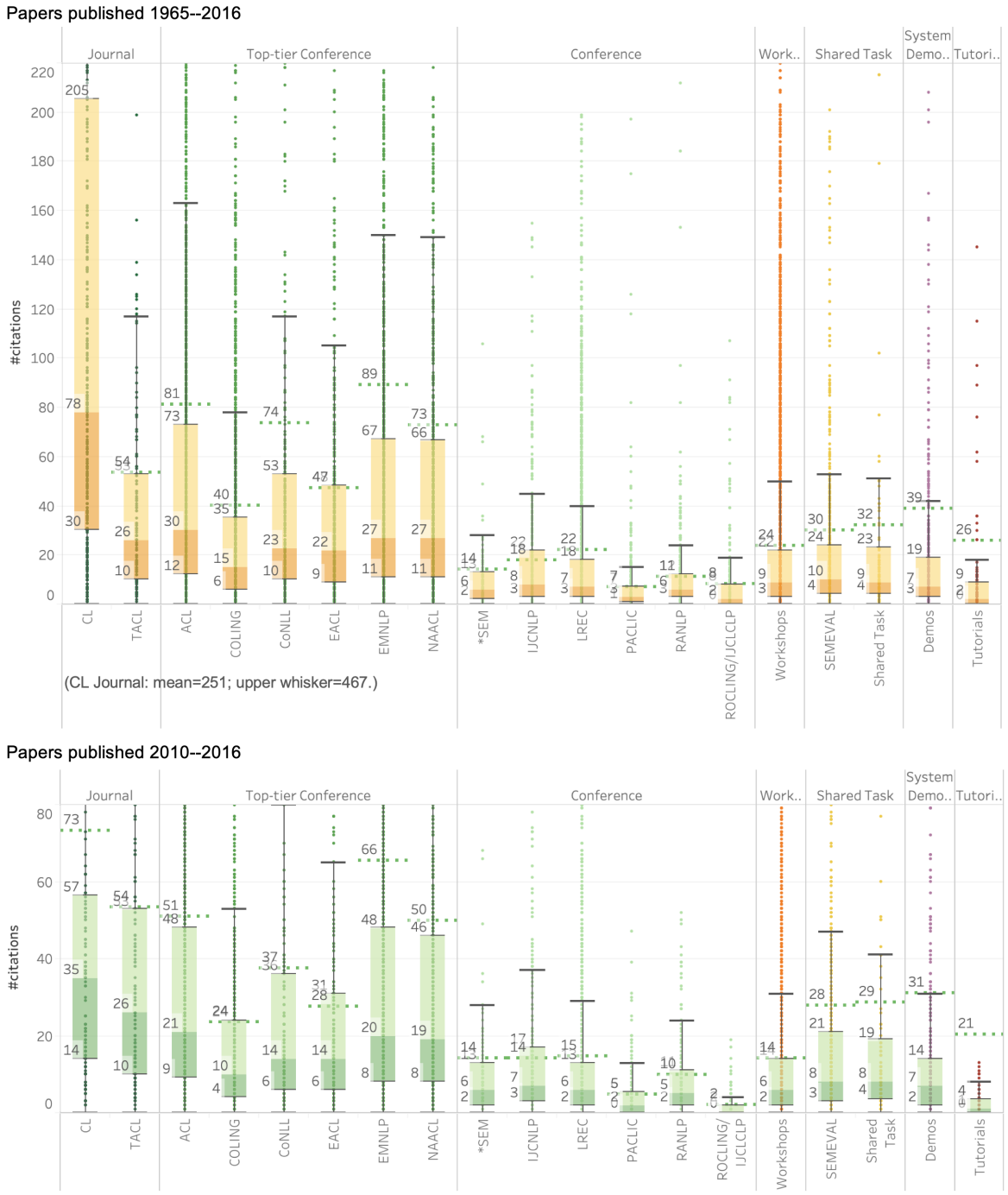

Figure 5: Citation box plots for papers by venue, type: papers published 1965–2016 (top) and papers published 2010–2016 (bottom).

Discussion: CL Journal has the highest median and average citation numbers. ACL comes second, closely followed by EMNLP and NAACL. The gap between CL Journal and ACL is considerably reduced when considering the 2010–2016 papers. IJCNLP and LREC have the highest numbers among the non-top-tier conferences, but their numbers remain lower than the numbers for SemEval, non-SemEval shared tasks, and workshops.

TACL, a journal, has substantially lower citation numbers than CL Journal, ACL, EMNLP, and NAACL (Figure 5 top). However, it should be noted that TACL only began publishing since 2013. (Also, with a page limit of about ten, TACL papers are arguably more akin to conference papers than journal papers.) When considering only the 2010–2016 papers, TACL's citation numbers are second only to CL Journal (Figure 5 bottom).

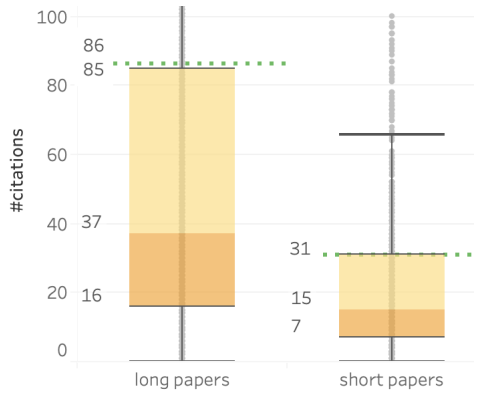

Figure 6: Citations box plots for long and short ACL papers published between 2003 and 2016.

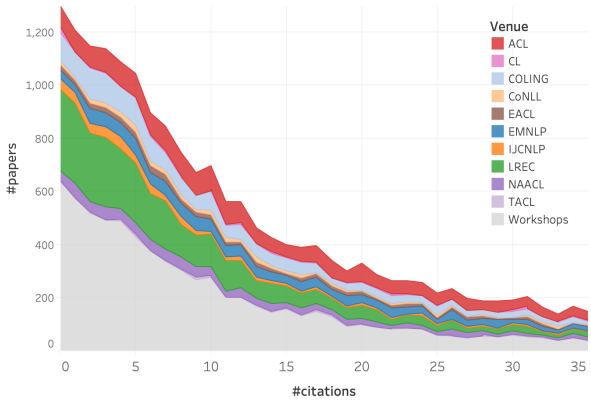

Figure 7: Stream graph of #papers by #citations. The contribution of each venue and paper type is stacked one on top of another.

When considering 2010–2016 papers, the system demonstration papers, the SemEval shared task papers, and non-SemEval shared task papers have notably high averages (surpassing or equalling those of COLING and EACL); however their median citations are lower. (This is consistent with the trends we saw earlier in Q3.)

Q6. How well cited are long and short ACL main conference papers, respectively?

A. Short papers were introduced by ACL in 2003. Since then ACL is by far the venue with the highest number of short papers (compared to other venues). So we compare long and short papers published at ACL since 2003 to determine their average citations. Figure 6 shows the citation box plots for long and short papers published between 2003 and 2016 at ACL. The two distributions are statistically different (Kolmogorov–Smirnov test, $p < .01$).

Discussion: In 2003, the idea of short papers was a novelty. It was conceived with the idea that there needs to a be a place for focused contributions that do not require as much space as a long paper. The format gained popularity quickly, and short papers at ACL tend to be incredibly competitive (sometimes having a lower acceptance rate than long papers). While there have been several influential short papers, it remains unclear how well-cited they are as a category. This analysis sheds some light to that end. We find that, on average, long papers get almost three times as many citations as short papers; the median for long papers is two-and-half times that of short papers.

Q7. How do different venues and paper types compare in terms of the volume of papers pertaining to various amounts of citation?

A. Figure 7 shows a stream graph of #papers by #citations. The contributions of each of the venues and paper types are stacked one on top of another (bands of colors). For a given point on the citations axis (say k), the width of the stream corresponds to the number of papers with k citations.

Discussion: It is not surprising to see that the #papers by #citations curve follows a power law distribution. (There are lots of papers with 0 or few citations, but the number drops of exponentially with the number of citations.) Workshop papers (light grey) are the most numerous, followed by LREC (green)—as observable from their wide bands. The bands for ACL, COLING, EMNLP, and NAACL are easily discernable but the bands for many others, especially CL Journal and TACL are barely discernable indicating low relative volume of their papers.

Observe that the bands for workshops and LREC are markedly wider in the 0 to 10 citations range than in the 11 and more citations range of the x axis. In contrast, the widths of the bands for top-tier conferences, such as ACL and EMNLP, remain relatively stable. Nonetheless, in terms of raw volume, it is worth noting that the workshops and LREC each produce more papers that are cited ten or more times than any other venue. As one considers even higher citations, the top-tier conferences become more dominant.

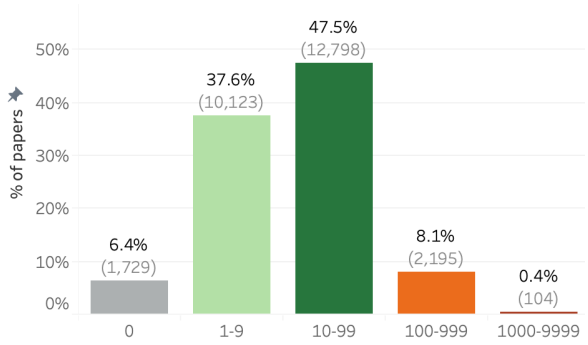

Figure 8: The percentage of AA′ papers in various citation bins. In parenthesis: #papers.

Q8. What percentage of papers are cited more than 10 times?[11] *How many papers are cited 0 times?*

A. Figure 8 shows the percentage of AA′ papers in various citation bins: 0, 1–9, 10–99, and 1000–9999. (The number of papers is shown in parenthesis.)

Discussion: About 56% of the papers are cited ten or more times. 6.4% of the papers are never cited. (Note also that some portion of the 1–9 bin likely includes papers that only received self-citations.) It would be interesting to compare these numbers with those in other fields such as medical sciences, physics, linguistics, machine learning, and psychology.

Q9. How well cited are areas within NLP?

A. We used word bigrams in the titles of papers to sample papers from various areas.[12] The title has a privileged position in a paper. It serves many functions, but most importantly, it conveys what the paper is about. For example, a paper with the bigram *machine translation* in the title is likely about machine translation (MT). We removed function words from the titles of papers in AA, and extracted all bigrams. Figure 9 shows, in order of decreasing frequency, the list of 66 bigrams that occurred in more than 100 papers. For each bigram, the yellow/green bar shows the median citations of the corresponding papers. The average citations and the number of papers are shown in parenthesis.

[11] Google Scholar invented the i-10 index as another measure of author research impact. It stands for the number of papers by an author that received ten or more citations. (Ten here is somewhat arbitrary, but reasonable.)

[12] Other approaches such as clustering are also reasonable; however, results with those might not be easily reproducible. We chose the title bigrams approach for its simplicity.

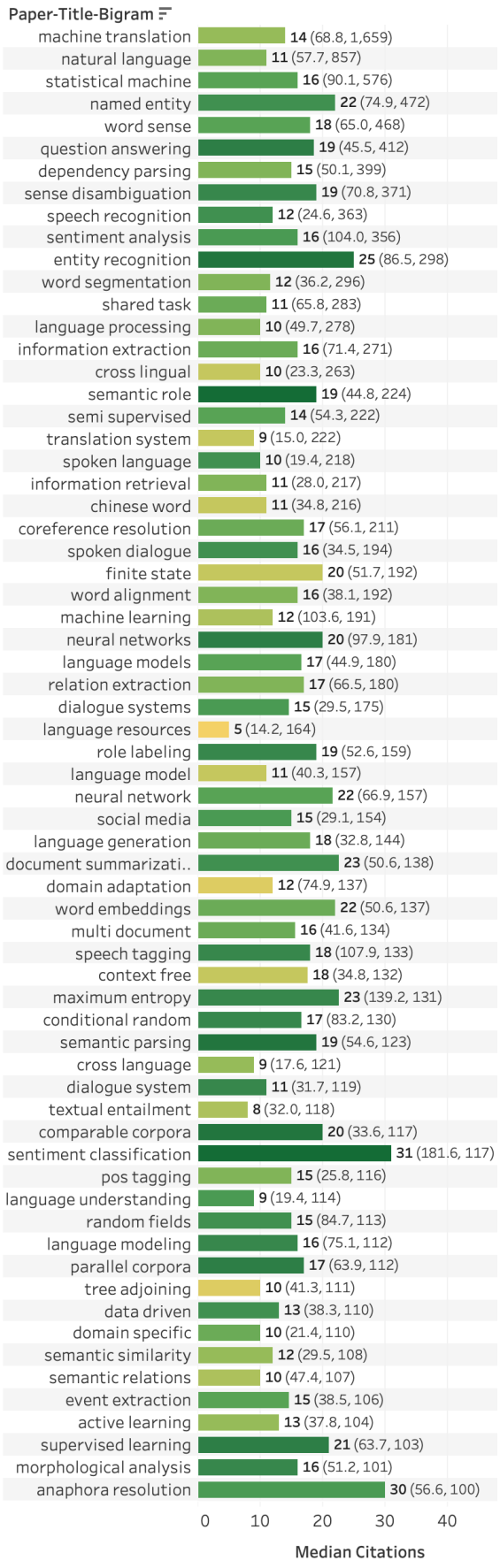

Figure 9: Bar graph of median citations. Title bigrams ordered by number of papers. In parenthesis: average citations, #papers.

Discussion: The graph shows, for example, that the bigram *machine translation* occurred in 1,659 AA′ papers that have a median citation count of 14, while the average is 68.8. The average is one of the highest among the bigrams, despite the median being more middle of the pack. This suggests the presence of heavily cited, outlier, papers. Indeed, the most cited paper in all of AA′ is an MT paper with more than 9000 citations (Papineni et al., 2002). Note that not all MT papers have *machine translation* in the title. Although non-random, this sample of 1,659 papers is arguably a reasonably representative sample of MT papers.

Third in the list are papers with *statistical machine* in the title—most commonly from the phrase *statistical machine translation*. One expects considerable overlap across these sets of papers. However, machine translation likely covers a broader range of research including work done before statistical MT was introduced, as well as work on neural MT and MT evaluation.

The bigrams with the highest median include: *sentiment classification* (31), *anaphora resolution* (30), and *entity recognition* (25). The bigrams with the lowest median include: *language resources* (5), *textual entailment* (8), *translation system* (9), and *cross language* (9). The bigrams with the highest average include: *sentiment classification* (181.6), *speech tagging* (107.9), *sentiment analysis* (104.0), and *statistical machine* (90.1).[13] One can access the lists of highly cited papers, pertaining to each of the bigrams, through the interactive visualization.

5 Limitations and Future Work

We list below some ideas of future work that we did not explore in this paper:

- Analyze NLP papers that are published outside of the ACL Anthology.
- Measure involvement of the industry in NLP publications over time.
- Measure the impact of research publications in other ways beyond citations. Identify papers that have made substantial contributions in non-standard ways.

A list of limitations and ethical considerations associated with this work is available online.[14]

[13]Note that simply composing titles with these high-citation bigrams is not expected to attract a large number of citations.

[14]https://medium.com/@nlpscholar/about-nlp-scholar-62cb3b0f4488

6 Conclusions

We extracted citation information for ∼1.1M papers from Google Scholar profiles of researchers who published at least three papers in the ACL Anthology. We used the citation counts of a subset (∼27K papers) to examine patterns of citation across paper types, venues, over time, and across areas of research within NLP.

We showed that only about 56% of the papers are cited ten or more times. CL Journal has the most cited papers, but the citation gap between CL journal and top-tier conferences has reduced in recent years. On average, long papers get almost three times as many citations as short papers. In case of popular shared tasks, the task-description papers and competition-winning system-description papers often receive a considerable number of citations. So much so that the average number of citations for the shared task papers is higher than the average for non-top-tier conferences. The papers on *sentiment classification*, *anaphora resolution*, and *entity recognition* have the highest median citations. Workshop papers and the shared task papers have higher median and average citations than the non-top-tier conferences.

The analyses presented here, and the associated dataset of papers mapped to citations, have a number of uses including, understanding how the field is growing and quantifying the impact of different types of papers. In separate work, we explored the use of the dataset to detect gender disparities in authorship and citations (Mohammad, 2020a). The dataset can potentially also be used to compare patterns of citations in NLP with those in other fields. Finally, we note again that citations are not an accurate reflection of the quality or importance of individual pieces of work. A crucial direction of future work is to develop richer ways of capturing scholarly impact.

Acknowledgments

This work was possible due to the helpful discussion and encouragement from a number of awesome people including: Dan Jurafsky, Tara Small, Michael Strube, Cyril Goutte, Eric Joanis, Matt Post, Patrick Littell, Torsten Zesch, Ellen Riloff, Iryna Gurevych, Rebecca Knowles, Isar Nejadgholi, and Peter Turney. Also, a big thanks to the ACL Anthology and Google Scholar Teams for creating and maintaining wonderful resources.

References

Ashton Anderson, Dan McFarland, and Dan Jurafsky. 2012. Towards a computational history of the acl: 1980-2008. In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pages 13–21.

Selcuk Aya, Carl Lagoze, and Thorsten Joachims. 2005. Citation classification and its applications. In *Knowledge Management: Nurturing Culture, Innovation, and Technology*, pages 287–298. World Scientific.

Steven Bird, Robert Dale, Bonnie Dorr, Bryan Gibson, Mark Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir Radev, and Yee Fan Tan. 2008. The ACL anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Lutz Bornmann and Hans-Dieter Daniel. 2009. The state of h index research. *EMBO reports*, 10(1):2–6.

Arthur R Bos and Sandrine Nitza. 2019. Interdisciplinary comparison of scientific impact of publications using the citation-ratio. *Data Science Journal*, 18(1).

Zoe Bulaitis. 2017. Measuring impact in the humanities: Learning from accountability and economics in a contemporary history of cultural value. *Palgrave Communications*, 3(1):7.

Daniel Gildea, Min-Yen Kan, Nitin Madnani, Christoph Teichmann, and Martín Villalba. 2018. The ACL anthology: Current state and future directions. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 23–28, Melbourne, Australia. Association for Computational Linguistics.

Michael Gusenbauer. 2019. Google scholar to overshadow them all? comparing the sizes of 12 academic search engines and bibliographic databases. *Scientometrics*, 118(1):177–214.

Jared L. Howland, Thomas C. Wright, Rebecca A. Boughan, and Brian C. Roberts. 2009. How scholarly is google scholar? a comparison to library databases. *College & Research Libraries*, 70(3).

John PA Ioannidis, Jeroen Baas, Richard Klavans, and Kevin W Boyack. 2019. A standardized citation metrics author database annotated for scientific field. *PLoS biology*, 17(8):e3000384.

Madian Khabsa and C Lee Giles. 2014. The number of scholarly documents on the public web. *PloS one*, 9(5):e93949.

Joseph Mariani, Gil Francopoulo, and Patrick Paroubek. 2018. The nlp4nlp corpus (i): 50 years of publication, collaboration and citation in speech and language processing. *Frontiers in Research Metrics and Analytics*, 3:36.

Alberto Martín-Martín, Enrique Orduna-Malea, Mike Thelwall, and Emilio Delgado López-Cózar. 2018. Google scholar, web of science, and scopus: A systematic comparison of citations in 252 subject categories. *Journal of Informetrics*, 12(4):1160–1177.

John Mingers and Loet Leydesdorff. 2015. A review of theory and practice in scientometrics. *European journal of operational research*, 246(1):1–19.

Saif Mohammad, Bonnie Dorr, Melissa Egan, Ahmed Hassan, Pradeep Muthukrishan, Vahed Qazvinian, Dragomir Radev, and David Zajic. 2009. 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*, pages 584–592.

Saif M. Mohammad. 2019. The state of nlp literature: A diachronic analysis of the acl anthology. *arXiv preprint arXiv:1911.03562*.

Saif M. Mohammad. 2020a. Gender gap in natural language processing research: Disparities in authorship and citations. In *Proceedings of the 2020 Annual Conference of the Association for Computational Linguistics*, Seattle, USA.

Saif M. Mohammad. 2020b. Nlp scholar: A dataset for examining the state of nlp research. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC-2020)*, Marseille, France.

Saif M. Mohammad. 2020c. Nlp scholar: An interactive visual explorer for natural language processing literature. In *Proceedings of the 2020 Annual Conference of the Association for Computational Linguistics*, Seattle, USA.

Hidetsugu Nanba, Noriko Kando, and Manabu Okumura. 2011. Classification of research papers using citation links and citation types: Towards automatic review article generation. *Advances in Classification Research Online*, 11(1):117–134.

Enrique Orduña-Malea, Juan Manuel Ayllón, Alberto Martín-Martín, and Emilio Delgado López-Cózar. 2014. About the size of google scholar: playing the numbers. *arXiv preprint arXiv:1407.6239*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318.

Son Bao Pham and Achim Hoffmann. 2003. A new approach for scientific citation classification using cue phrases. In *Australasian Joint Conference on Artificial Intelligence*, pages 759–771. Springer.

Jason Priem and Bradely H Hemminger. 2010. Scientometrics 2.0: New metrics of scholarly impact on the social web. *First monday*, 15(7).

Vahed Qazvinian, Dragomir R Radev, Saif M Mohammad, Bonnie Dorr, David Zajic, Michael Whidby, and Taesun Moon. 2013. Generating extractive summaries of scientific paradigms. *Journal of Artificial Intelligence Research*, 46:165–201.

Dragomir R Radev, Mark Thomas Joseph, Bryan Gibson, and Pradeep Muthukrishnan. 2016. A bibliometric and network analysis of the field of computational linguistics. *Journal of the Association for Information Science and Technology*, 67(3):683–706.

James Ravenscroft, Maria Liakata, Amanda Clare, and Daniel Duma. 2017. Measuring scientific impact beyond academia: An assessment of existing impact metrics and proposed improvements. *PloS one*, 12(3):e0173152.

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006. Automatic classification of citation function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 103–110.

Alan M Turing. 1950. Computing machinery and intelligence-am turing. *Mind*, 59(236):433.

Alan M Turing. 2009. Computing machinery and intelligence. In *Parsing the Turing Test*, pages 23–65. Springer.

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

Xiaodan Zhu, Peter Turney, Daniel Lemire, and André Vellino. 2015. Measuring academic influence: Not all citations are equal. *Journal of the Association for Information Science and Technology*, 66(2):408–427.