Using Named Entity Recognition as a Classification Heuristic

Andrea K. Thomer[1] and Nicholas M. Weber[1]

1 Center for Informatics Research in Science and Scholarship, Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign

Abstract

This poster proposes the use of Named Entity Recognition as a heuristic tool for improving manual document classification. This technique was developed as part of a project studying collaborative work via the acknowledgment statements found in a corpus of formally published journal articles. We demonstrate how uncertainty in our initial text mining results were 'ground-truthed' using Natural Language Processing tools in a quick-and-dirty fashion. To verify this technique's validity, we offer some initial results from our larger study.

Keywords: bioinformatics, text mining, natural language processing, named entity recognition, acknowledgments, authorship

Citation: Thomer, A. K., & Weber, N. M. (2014). Using Named Entity Recognition as a Classification Heuristic. In *iConference 2014 Proceedings* (p. 1133–1138). doi:10.9776/14401

Acknowledgements: Thanks to the anonymous reviewers who provided us with excellent and helpful feedback. Thanks to the makers of the Stanford Named Entity Recognizer for making their tools openly available.

Contact: thomer2@illinois.edu, nmweber@illinois.edu

1 Introduction

The formally published scientific journal article has been mined, examined and evaluated in nearly every aspect; titles, authorship lists, abstracts, methods, figures, footnotes, and citations have all been used to better understand the way a field of science communicates, collaborates and makes new knowledge claims.

Past work has shown that the "acknowledgments" section of a journal article can be especially helpful in shedding light on the often neglected, or invisible work of collaboration (Cronin, Shaw and Labarre, 2003; 2004), especially in domains that depend on expert methodological knowledge and instrument building (Salager-Meyer et al, 2010). As part of an on-going research project, we're exploring acknowledgment statements found in a large corpus of bioinformatics texts to better understand collaborations between the diverse peoples, technologies, and research tools that produce computational biological knowledge. In particular, we want to better understand how successful interdisciplinary collaborative arrangements distribute credit, how material resources are cited, and how computational and biological knowledge have subtly blended in this field over time. In a field like bioinformatics, research questions about acknowledgment and authorship practices are further complicated by the increased scale of collaboration, and the heterogeneity of scholarly products generated over the course of a research project (e.g. code, datasets, executable workflows) which are not easily attributable to one, or even a few "authors." Understanding how credit is established and formally recognized in this field will help policy makers better understand and design incentives and reward structures so that both funding agencies and information systems developers might optimize cooperative work arrangements (Howison and Herbsleb, 2011; 2013).

Our work diverges from previous studies of acknowledgment in some important methodological ways. Past studies relied upon the manual extraction of bibliographic data, and the labor-intensive annotation of acknowledgment texts for the purposes of later classification (Giles and Councill, 2004 a notable exception). Here we present our first steps towards applying natural language processing (NLP) techniques, as well as text mining methods to extract acknowledgment texts from a corpus of documents gathered from the PubMed Central Open Access collection. During this phase of research we have focused

on finding economic ways to increase the speed of our classifications without sacrificing accuracy, nor reliability. In that vein, our research questions include the following:

- With little to no customization, can NLP tools like the Stanford Named Entity Recognizer (Stanford NER) help us initially evaluate the quality of a corpus of acknowledgment statements? And, can they identify "entity rich" acknowledgments on which we should focus our initial analysis?
- How effective are general, out-of-the-box NLP tools at recognizing entities in a domain specific corpus (such as bioinformatics)?
- How can we best leverage tools that deliver quantitative results (e.g. number of entities per acknowledgment statement) to support or aide further qualitative enquiry?

2 Methods

2.1 Corpus Construction

We assembled a representative collection of bioinformatics texts from PubMed Central's Open Access (PMC-OA) corpus. The PMC-OA includes the full text of completely open access journals, and the NIH-portfolios of other paid access journals. We selected texts from two high-impact, open access bioinformatics journals (*PLoS Computational Biology* (n=2776) and *BMC Bioinformatics* (n=5765)) and one high-impact, limited access journal (the NIH portfolio from *Bioinformatics* (n=1200)) (Table 1). Each article is encoded in .nxml format, utilizing Z39.96, the Journal Archive Tag Suite (JATS).

	Bioinformatics	BMCBioinformatics	PLoSComputBiol	Total
2000		1		1
2001		9		9
2002		40		40
2003		66		66
2004		209		209
2005		371	71	442
2006		633	169	802
2007	1	599	251	851
2008	144	731	298	1173
2009	269	729	394	1392
2010	279	845	422	1546
2011	201	719	426	1346
2012	242	601	530	1373
2013	64	212	215	491
Total	1200	5765	2776	9741

Table 1: All articles in the corpus were published between 2001-2013; n=9741.

2.2 Text-mining acknowledgments

Utilizing BeautifulSoup[1], a Python library that supports html and xml processing, we wrote a series of scripts to extract acknowledgments sections from each article[2]. Because of PMC-OA's's use of the JATS markup, extraction of these statements was straightforward for the majority of our sampled articles (5897),

1 http://www.crummy.com/software/BeautifulSoup/

2 code available at https://github.com/akthom/ParatextsAndDocumentaryPractices

which encoded their acknowledgment statements with the JATS *<ack>* tag, intended to specifically designate acknowledgment text.

We found that a large portion of the articles encoded their acknowledgment statements using a combination of the more general *<back>* and *<sec>* tags, which are catchalls for many of an article's back matter, and any discrete section of an article, respectively. Our more general script extracting the contents of both *<ack>* and *<back>* tags pulled an additional 2377 sections of text (total statement extracted: 8427, or 86.5% of the total corpus), with an estimated 1% error rate. We also extracted each article's author list, and tallied the total number of authors per article (see Figure 1).

2.3 Named Entity Recognition

After text mining the acknowledgment statements from our corpus of bioinformatics documents (n=9741) we parsed the texts with the Stanford Named Entity Recognizer (Stanford NER; Finkel, Grenager & Manning, 2005) using a 4 class model trained to recognize and tag persons, organizations, locations and miscellaneous "other" entities. We then manually reviewed a small random sample of the results (n=100) to review the NER's efficacy.

3 Results

Overall, the Stanford NER identified 21985 unique persons, 30223 Organizations, 10444 Locations, and 5423 Misc entities. After manually reviewing results from a sample of acknowledgment statements we found that the *person* entity tagger was by far the most accurate, and helped us further explore whom was acknowledged, and how often. While the *organization* tagger worked fairly well (with over 60% accuracy in our reviewed sample), it would sometimes parse organizations with compound names into more than one entity (e.g. "Center for <ORGANIZATION>Insect Science</ORGANIZATION> at the <ORGANIZATION>University of Arizona</ORGANIZATION>). *Misc* entities proved unreliable, and too difficult to assess (the Stanford NER often erroneously tagging adjectives like "Open Access" and "Dutch" as entities, while also tagging entities that could arguably be classified as organizations, such as the "OBO Edit Working Group"). We do, however, note that the *misc* tagger did identify a number of computing facilities and software packages as entities, giving us hope that the method could be altered to automatically extract computational entities in the future.

We compiled a list of the most commonly acknowledged persons in our corpus, and then tried to identify each person's title and institutional affiliations using author affiliations from the articles themselves, and then generic internet searches to further flesh out each person's role within an institution (Table 2).

Name	# ack	Job title
Elena Rivas	16	Janelia Senior Scientist, Howard Hughes Medical Center*
Vasant Honavar	11	Professor of Computer Science and head of Artificial Intelligence Research Lab, Iowa State University*
Burkhard Rost	10	Computational Biologist and Computer Scientist, Technical University of Munich*
Chris Mungall	10	Bioinformatics Scientist, Lawrence Berkeley National Lab
Gary Bader	10	Professor of Molecular Genetics and Computer Science, The Donnelly Centre, University of Toronto*
Terry Mark-Major	10	Business Manager, University of Tennessee Health Science Center
Alex Skrenchuk	9	IT Manager, Stanford Center for Biomedical Informatics Research

Alexander Zien	9	Research Scientist, Max Planck Institute for Intelligent Systems
Eran Segal	9	Professor and Computational biologist, Weizmann Institute of Science*
Isobel Peters	9	Senior Project Manager, BioMed Central
		* appears to manage her/his own lab

Table 2: The ten most frequently acknowledged individuals in our corpus.

We found that the ten most frequently acknowledged individuals were evenly split between researchers who are the director or lead scientist of a lab, and researchers who appeared to have support staff roles. In this case, NER-augmented classification helped us quickly see that our dataset contained information relevant to our broader research questions regarding the invisible work of collaborative projects, and encouraged us to further explore the relationship between authorship and acknowledgment within this corpus.

We compared the number of authors per article per year to the number of acknowledged individuals per article per year, to get a sense of whether there were any noticeable authorship or acknowledgment trends within bioinformatics publications more generally (Figure 1).

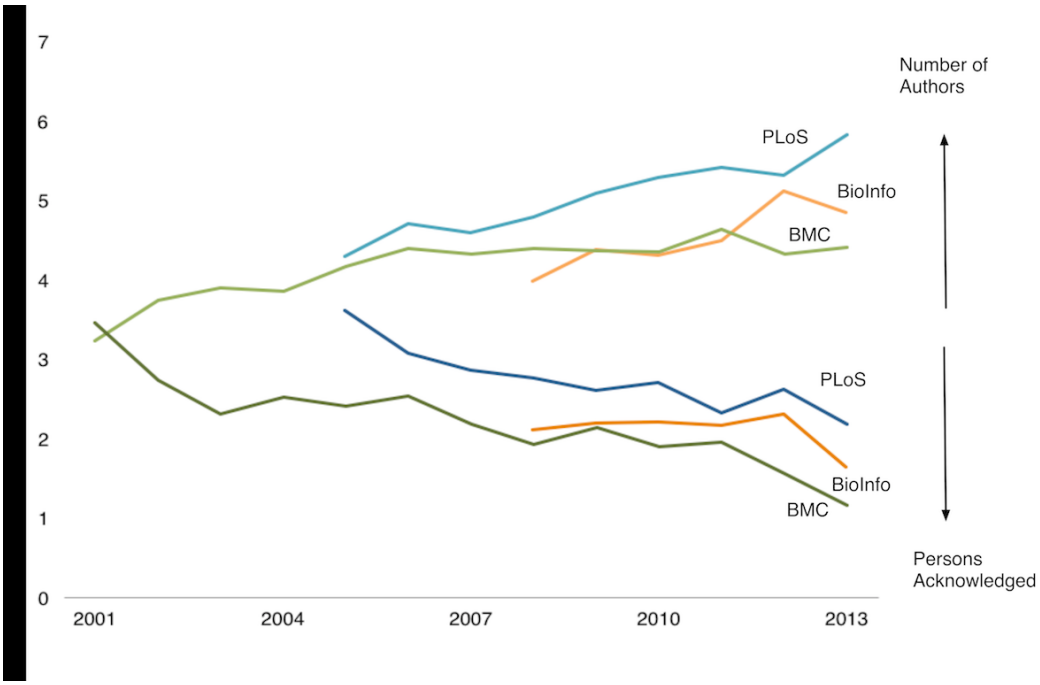

Figure 1: Average number of authors per article per year compared to the average number of acknowledged individuals per article per year.

Interestingly, we noted slight downward trends in the number of acknowledged individuals per article per year, apparently corresponding with slight upward trends in the number of authors per article per year. One possible explanation for this trend is that the *BMC Bioinformatics* and *PLoS Computational Biology* collections both include editorial matter in addition to peer reviewed journal articles, and the *PLoS* corpus also includes conferences proceedings; thus the downward trends in number of acknowledged persons per article could be the result of increased inclusion of articles without acknowledgments sections thereby "watering down" our results and making it appear as if the number of acknowledged individuals is decreasing.

This has encouraged us to look at differences between types of publications and whom, or what, was acknowledged; our future work will explore how acknowledgment and authorship differ between regular publications, software publications (somewhat unique to bioinformatics publishing) and conference proceedings. Using NER as a rough classification heuristic allowed us to narrow in on this area relatively quickly, and sensitized us to the relationship for future work.

4 Conclusions and next steps

We have found that using NLP tools in a heuristic way can be quite helpful in quickly evaluating the relevance of a corpus for further, more rigorous analysis – and furthermore, for identifying future directions in the development of named entity recognizers. In the context of our larger project, use of NER tools helped us quickly determine the relevance of bioinformatics acknowledgment statements to studies of collaboration, and to determine whether or not the number and types of named entities would warrant further manual classification.

This quick and dirty work encouraged us to continue analyzing our named entities in conjunction with our manual classification of acknowledgment types and tropes. It also helped us recognize the important relationship between acknowledgments and authorship statements. In future work we hope to apply our methods to a more diverse corpus of acknowledgment statements, to further explore underlying reasons for the above trends in authorship and acknowledgment rates, and to examine the relationship between article type, editorial policy, and acknowledgment practices. Additionally, we hope to explore customization of a named entity recognizer specific to the needs of this work; an NER designed to identify computing facilities and software would not only aid us in our research, but could also more generally support scientometric analysis of the impact of computational resources.

Finally, we note that named entity recognition may provide publishers and researchers alike with a way to augment existing text encoding schemas, such as JATS. While the JATS markup facilitates more precise entity extraction, it is unrealistic to expect publishers (and text encoding schema developers) to encode all possible entities of interest. *Post hoc* named entity extraction can supplement metadata-facilitated information extraction efforts, particularly in fields like bioinformatics, in which authorship and acknowledgment practices may be rapidly evolving.

5 References

Cronin, B., Shaw, D., & Labarre, K. (2003). A cast of thousands: Coauthorship and subauthorship collaboration in the 20th century as manifested in the scholarly journal literature of psychology and philosophy. *Journal of the American Society for Information Science and Technology*, *54*(9), 855-871.

Cronin, B., Shaw, D., & Labarre, K. (2004). Visible, less visible, and invisible work: Patterns of collaboration in 20th century chemistry. *Journal of the American Society for Information Science and Technology*, *55*(2), 160-168.

Finkel, J., Grenager, T., & Manning, C. (2005). Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005),* pp. 363-370. http://nlp.stanford.edu/~manning/papers/gibbscrf3.pdf

Giles, C. L., & Councill, I. G. (2004). Who gets acknowledged: Measuring scientific contributions through automatic acknowledgment indexing. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(51), 17599-17604.

Howison, J., & Herbsleb, J. D. (2011). Scientific software production: incentives and collaboration. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work* (pp. 513-522). China : ACM.

Howison, J., & Herbsleb, J. D. (2013, February). Incentives and integration in scientific software production. In *Proceedings of the 2013 conference on Computer supported cooperative work* (pp. 459-470). Texas: ACM.

Salager-Meyer, F., Ariza, M. Á. A., & Berbesí, M. P. (2009). "Backstage solidarity" in Spanish-and English-written medical research papers: Publication context and the acknowledgment paratext. *Journal of the American Society for Information Science and Technology*, *60*(2), 307-317.

6 Table of Figures

7 Table of Tables