Discovery Informatics: AI Takes a Science-Centered View on Big Data
AAAI Technical Report FS-13-01

Finna: A Paragraph Prioritization System for Biocuration in the Neurosciences

Kyle H. Ambert
Intel Labs, Graph Analytics Operation
Hillsboro, OR

Aaron M. Cohen
Oregon Health & Science University
Portland, OR

Gully APC Burns
University of Southern California
Los Angeles, CA

Eilis Boudreau
Oregon Health & Science University
Portland, OR

Kemal Sonmez
Oregon Health & Science University
Portland, OR

Abstract

The emphasis of multilevel modeling techniques in the neurosciences has led to an increased need for large-scale, computationally-accessible databases containing neuroscientific data. Despite this, such databases are not being populated at a rate commensurate with their demand amongst Neuroinformaticians. The reasons for this are common to scientific database curation in general, namely, limitation of resources. Much of neuroscience's long tradition of research has been documented in computationally inaccessible formats, such as the *pdf*, making large-scale data extraction laborious and expensive. Here, we present a system for alleviating one bottleneck in the workflow for curating a typical knowledge base of neuroscience-related information. *Finna* is designed to rank-order the composite paragraphs of a publication that is predicted to contain information relevant to a knowledge base, in terms of the probability that each documents relevant data. We were able to achieve excellent performance with our classifier (AUC > 0.90) on our manually-curated neuroscience document corpus. Our approach would allow curators to read only a median of 2 paragraphs for each document, in order to identify information relevant to a neuron-related knowledge base. To our knowledge, this is the first system of its kind, and will be a useful baseline for developing similar resources for the neurosciences, and curation in general.

Introduction

The manual creation of discipline-specific knowledge bases is an expensive and time-consuming process that requires the efforts of domain experts over an extended period of time. Although some general-purpose tools have been developed for streamlining biocuration workflows (Burns et al. 2009; Hirschman et al. 2012; Karamanis et al. 2007; Pokkunuri et al. 2011; Ramakrishnan et al. ; Rodriguez-Esteban, Iossifov, and Rzhetsky 2006; Wiegers et al. 2009; Wiegers 2009; Burge et al. 2012), and some work has been directed toward developing task-specific solutions using text-mining (particularly for the curation of systematic reviews (Ananiadou et al. 2009; Cohen, Ambert, and McDonagh 2009; 2009; Cohen et al. 2010; 2010; Cohen, Ambert, and McDonagh 2010; Wallace et al. 2010; Yang, Cohen, and McDonagh 2008)), there remain time-consuming tasks in biocurators' workflows that can be made more efficient using machine learning methods. Many of these approaches have focused on classifying documents in terms of whether they contain information that is likely to be relevant to the curation task at hand. This, however, only solves one part of the efficiency problem–given a likely relevant document, curators must still spend time reading through it, in an effort to locate the information that will eventually be included in the knowledge base.

Here, we describe *Finna*, a system which, given a document that is likely to contain information relevant to a scientific publication-sourced knowledge base that is narrow in scope, will re-order its composite paragraphs in terms of the likelihood that each contains the information of interest. This addresses an important bottle-neck in typical biocuration workflows (Ambert and Cohen 2012). As a part of the curation process, knowledge base curators review any submissions they believe necessitate review (either from manual prioritization of submissions, or using automated machine learning-based approaches, as in (Cohen, Ambert, and McDonagh 2010; Ambert and Cohen 2012; Yang, Cohen, and McDonagh 2008)). Once a potentially-relevant publication has been identified, curators typically need to manually extract the information, reading most or all of the publication to do so. The goal of our system is to minimize the amount of time curators need to spend reading information that will not, ultimately, contribute to, and speed up, the development of their knowledge base, thus helping them efficiently use their time.

Our work here uses the entries in the Neuron Registry (NR), a community-curated knowledge base under the direction of the Neuron Registry Task Force (NRTF[1]), a part of the International Neuroinformatics Coordinating Facility (INCF) Program on Ontologies of Neural Structures (PONS). The primary goal of the NRTF is to create the infrastructure for machine-readable knowledge bases of neuronal cell types,

[1] http://pons.neurocommons.org/page/Neuron_registry

providing a formal means for describing and quantifying existing cell types from their properties, and populating it with information that has been extracted from the primary literature. The Neuron Registry, in particular, was chosen for our work for several reasons. First, biologically-sound multilevel models of neural circuitry will all necessarily leverage the information contained in a database of neuronal attributes. As such, the Neuron Registry stands to become an important source of modeling information in Neuroinformatics. Second, is the small amount of already curated information contained in the Neuron Registry. The process of identifying and verifying new material for inclusion in the such a database is laborious, and so it is unsurprising that the Neuron Registry houses only a small amount of information to date, which means it will benefit from our efforts. Finally, our previous work in helping bootstrap the development of the NR (Ambert 2013) means that there are sufficient positive-class publications referenced in the knowledge base to adequately build a baseline system of the sort we are interested in here.

Methods

Constructing the Paragraph Document Collection

The full text of all positive-class documents with a PubMed ID that were added to the NR during 2012 were acquired and randomly assigned to either a training set (128 documents), or a hold-out testing set (33 documents). Every document in the entire corpus was broken up into its respective composite paragraphs, following plain-text extraction, using the *pdftotext* utility (a part of the xpdf tool suite[2]), and cleaning up, with a set of regular expressions. Paragraphs were inferred, based on new-line separation–abstract sections and titles typically constituted their own paragraphs. Each NR entry was located in the text of its associated reference by a trained Neuroscientist (KHA), and the section name and paragraph number in which each was found were recorded. Thus, the paragraphs contained in every document were assigned either a positive- or negative-class label, according to whether each contained the NR information associated with their respective documents. After this procedure, the training collection had 9983 paragraphs (158 positive-class, 9825 negative-class), and the testing collection had 2026 paragraphs (35 positive-class, and 1991 negative-class).

Paragraph Classifier System Design

Of the many text classification algorithms available to use, we limited our system design experiments to Support Vector Machine (SVM) algorithm (linear kernel, c-parameter set to 1.0) (Joachims 1998), as our previous work has shown it to perform well on neuroscience-related text (Ambert and Cohen 2012; Ambert 2013), and biomedical text in general (Ambert and Cohen 2009; Cohen 2008).

[2]http://www.foolabs.com/xpdf/

Feature Methods The present task differs from typical document classification tasks in an important way: many of the typical sorts of features that one would use (e.g., document title, abstract, MeSH terms) can't be used here, since that type of document-level metadata is associated with every paragraph in a given document. Thus, we constructed a set of features that apply specifically to only a single paragraph at a time–n-grams of varying lengths ($n = 1-5$), and a set of quantitative features derived from the text contained in a paragraph. We used the vertebrate neuron branch of the community-curated ontology NeuroLex[3] to create regular expressions for identifying the presence of each neurolex term within the text of the paragraphs. Initial investigation showed that the only term which was directly identifiable in the training corpus was *retinal ganglion cell*, which occurred multiple times in both positive- and negative-class documents. This is not entirely surprising–neuroscience is a heterogeneous research discipline that is composed of investigators from a variety of academic backgrounds, and each has its own particular way of referring to neuroanatomical entities and writing conventions. For example, *amygdala basolateral nucleus stellate neuron*, could be referred to in a publication as *stellate neuron in the basolateral nucleus of the amygdala*, which would not be matched by our approach. Because manually creating a list of regular expressions covering all the possible ways each of the 247 neuroanatomical entities in the NeuroLex would be impractical, we instead hand-selected substrings from the NeuroLex entries which would likely be found in whatever way someone used to refer to the concept. For example, the *amygdala basolateral nucleus stellate neuron* became *stellate neuron*, and the *retinal ganglion cell* became ganglion cell. In addition, we hypothesized that annotated information might tend to occur in similar places between documents. Thus, we created a continuous-valued feature indicating the absolute paragraph order number within a document. Since recursive partitioning tended to be the most useful continuous-valued feature modeling technique in the previous study, we used it here as well. Here, our recursive partitioning method consisted of transforming the continuous-valued feature into a series of binary features representing sub-ranges of the features values in the training data; the cut-off points for these sub-ranges were selected based on the minimum description length principle (Fayyad and Irani 1993).

System Evaluation Since this the classification task described in this study is fundamentally a ranking task, we used AUC as our primary performance metric, which we defined as the area under the precision/recall curve. System configurations were done by performing five repetitions of 2-way cross-validation on the training data set. The feature configurations that showed some usefulness were then used in a train/classify experiment, in which a model is trained using the training data set, and classified using the hold-out test data set.

As a secondary performance metric for our train/classify experiments, we examined the median number of paragraphs

[3]http://neurolex.org/wiki/Vertebrate_Neuron_overview

that would need to be read for each publication in the hold-out classification set, in order to identify a paragraph containing information that is relevant to the neuron registry. We re-examined each the best-performing system in the train/classify experiments from this perspective. For the purposes of interpreting this metric, we compared this value to the median number of paragraphs a reviewer would have to read, if they started from the first paragraph, as published in the *pdf* version of the manuscript, and stopped once they reached the sentence(s) leading to the manuscript being included in the NR.

Results

Our initial classification experiment was done using five repetitions of two-way cross-validation on the training data collection. We examined two sets of system configurations–n-grams, where n=1-5, with paragraph locations, and paragraph locations alone. The results of this experiment are shown in Figure 1. From this figure, it is clear that increasing the number of n in the n-grams improves performance, which asymptotically approaches its peak by $n = 5$. Adding paragraph locations give a small, but not insignificant, performance boost, with maximum performance being achieved at 5-grams with paragraph location information.
We ran the same system configurations training on the entire training collection, and evaluating the resultant models against the hold-out testing collection. The results of these experiments are shown in Figure 2. The results observed here mirror those seen in the cross-validation experiments–a steady increase in performance with increasing size of n-grams, and a small performance boost obtained by adding paragraph location information. Interestingly, if we compare the performance increases between 4-gram and 5-gram in the cross-validation and train/classify experiments, it appears that AUC has begun to level off in the train/classify experiments, implying that investigating larger n-grams, or more training data, may not lead to dramatic improvements in performance. Based on these observations, we chose 4-grams with paragraph location information as our best-performing system, since it performed negligibly worse than the 5-gram system (0.906 v. 0.907), and was considerably faster.

Finally, we looked at the median number of paragraphs in the best-performing system (4-grams, and paragraph locations) that would have to be read in each document, in order to find the paragraph containing the positive-class sentence, if they were read in order of the system-generated rankings. We chose median as our measure of central tendency, as the distributions were highly positively skewed). We compared this value to the standard approach to document review–starting from the first paragraph, and reading until the information leading to a positive annotation has been identified. Based on looking at the distribution of paragraph rankings and paragraph locations (the standard) (see Figure 3), the proper measure of central tendency for these metrics appears to be the median. For the *Finna* approach, a median of 2 paragraphs in each document would have to be read by annotators, whereas in the standard approach, a median of 6 paragraphs would have to be read. The respective shapes of these distributions highlights another interesting difference. Comparing the distributions tails (e.g., from 20 paragraphs read and up), *Finna* tends to perform better–the standard approach has a consistently diminishing number of documents that require reading many paragraphs, as one moves up the x-axis, whereas there were only a handful where this would have been the case for the *Finna* system (e.g., at around 27 and 40).

The *Finna* system tended to perform very well on the majority of documents (1-4 paragraphs needing to be read), while the standard approach was more variable ($sd_{Finna} = 9.4$ v. $sd_{Standard} = 15.4$), resulting in many documents requiring over 20 paragraphs to be read. Although *Finna* yielded an average 7.4 paragraphs savings in reading, over the standard approach, there were a few documents that were outliers, in system performance. To delve into our results further, we examined Finna's performance in terms of the document locations of the annotated information (abstract, results, methods, & figure captions). Table 1 summarizes the results of this analysis. Overall, the *Finna* system tended to out-perform the standard approach, especially resulting in large reading savings for annotations found in the results section and figure captions. Annotations found in the document title, however, tended to lead to a small loss in paragraphs needing to be read.

Discussion

The present set of experiments describes a classification system that can be used for rank-ordering the paragraphs within a manuscript, in terms of their likelihood for containing information of interest to a neuroscience biocuration task. By training off of a subset of expert annotated documents known to contain information that is relevant to the NR, (Ambert 2013) we were able to create a system that performed well, both in terms of AUC (0.906), and a new metric, median number of paragraphs to read in a set of documents (median = 2).

We observed some variation in system performance, depending on where the annotatable information was found in the original document. In general, *Finna* lead to a savings in the amount of reading needing to be done by annotators, as compared with the standard approach, except in the case of information found in document titles. Although the standard approach is a logical standard of comparison, in terms of evaluating the practical utility of a paragraph re-ranking system, for understanding the performance from a machine learning perspective, this may not be the best metric. For example, using our paragraph parsing method, the title is always the first paragraph of the document, so any small changes in paragraph rank by our system can result in apparent performance degradation. Results sections, conversely, tend to occur toward the end of a manuscript, putting paragraph re-rankers at an advantage, from the perspective of evaluating in terms of the number of paragraphs reviewers have to read. Here, our system lead to a mean savings of 17.6 paragraphs, for documents with

Subsection	Count	Finna	Standard	Reading Savings (Finna)
Abstract	16	1.1	3.8	2.7
Title	3	5.0	2.0	-3.0
Results	7	11.1	28.7	17.6
Figure Captions	7	13.6	26.0	12.4
Methods	0	NA	NA	NA
Summary	33	2	6	7.4

Table 1: Table summarizing the results of the *Finna paragraphs to read* analysis, broken down by the mean results for the document section in which the annotated information was found. The *Finna* system tended to out-perform the standard approach, unless the annotated information was found in the Title section. Values in the summary row have different interpretations: Count column–count, Finna column–median performance, Standard column–median performance, Reading Savings column–mean savings. Data for annotations found in the Methods section are not present here, as, in the hold-out testing document collection, no annotatable information was found in the Methods section.

annotatable information in the results section. This is not to say that median paragraphs to read is a useless metric. Rather, systems such as *Finna*, that need to be optimized in terms of machine learning performance and practical performance alike should be evaluated using more than one metric. Here, we've created a system that performs well using standard machine learning metrics for document ranking (AUC), as well as in terms of a new metric designed to measure performance for how the system will actually be used.

An interesting question for future research to address is whether it is practical for curators to read a document's composite paragraphs out of order. Some curators may be uncomfortable reading paragraphs outside of the context in which the author intended them to be read; future work should compare interpretations of NR information based on presentation of the paragraph in which it was found, versus based on presentation of the document in which it was found. It would also be interesting to examine how a system which always presents documents' titles and abstracts first, followed by a prioritized list of the remaining paragraphs in the publication.

Based on our results, this system is ready for evaluation as a component of real-world biocuration workflow, likely working in consort with a document classification system for identifying publications that are likely to be relevant to a knowledge base (Ambert 2013). In this situation, biocurators would use a classification system to identify publications that are highly-likely to contain information of relevance to a particular biocuration task, and the system would automatically re-order the paragraphs of each predicted positive-class publication in terms of their likelihood for having the information that lead to being assigned a positive-class label. In order to be truly useful to a team of biocurators, such a system we need to be evaluated within the context of their workflow–it would need to be integrated in the least-obtrusive way, ideally allowing the curators to focus on adding expert-annotated data to their databases, rather than on using the system.

One limitation of our system is that the data it was trained on was generated by a single Neuroscientist. Although the curator had graduate-level training in the neurosciences, it is not possible at this time to evaluate the extent to which his annotations are generalizable to the larger population of neuroscience-related publications. To do this, a data set annotated by multiple experts would be needed, so that inter-annotator agreement statistics could be computed. Since the present set of experiments only used paragraph-level information for classification, one possible extension of the *Finna* system would be to incorporate document-level information into the classification workflow. For example, it may be possible to use the MEDLINE-derived MeSH terms to adjust the prior probability of certain paragraphs containing important information: the occurrence of certain NeuroLex entries (Bandrowski 2011) in the MeSH terms and certain paragraphs may imply that paragraph is more likely to succinctly communicate the main finding(s) of the manuscript. Beyond this, another way to improve future systems would be to model MEDLINE-related elements of the document (e.g., abstract and title) separately, removing them from the document. This would likely improve *Finna*'s handling of the Title paragraphs.

Conclusion

In this study we have demonstrated that an automated system can be used to rank-order a publication's paragraphs, in terms of their likelihood for containing information that is of interest to a biocuration task in the neurosciences. By training on a manually-curated neuroscience document set that has been pre-filtered to only include documents that do, in fact, contain information of interest, we showed that a simple configuration of an SVM-based classification system can be used to identify paragraphs of interest. We were able to achieve an AUC of 0.906 with our selected system, which was able to complete the classification task on the order of several seconds. Based on its level of performance and its speed, this system could be integrated into a real-world biocuration workflow, and used to streamline the process of curating a knowledge base.

Acknowledgments

We wish to acknowledge the contributions of Brian Roark

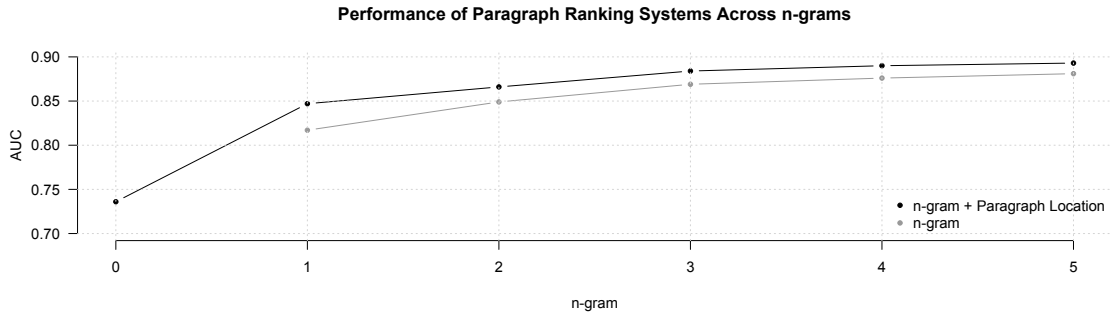

Figure 1: Performance of two configurations of the Finna system across increasing n-grams during five repetitions of two-way cross-validation. The n-gram with paragraph location system was better than n-grams alone, but there was not a great difference.

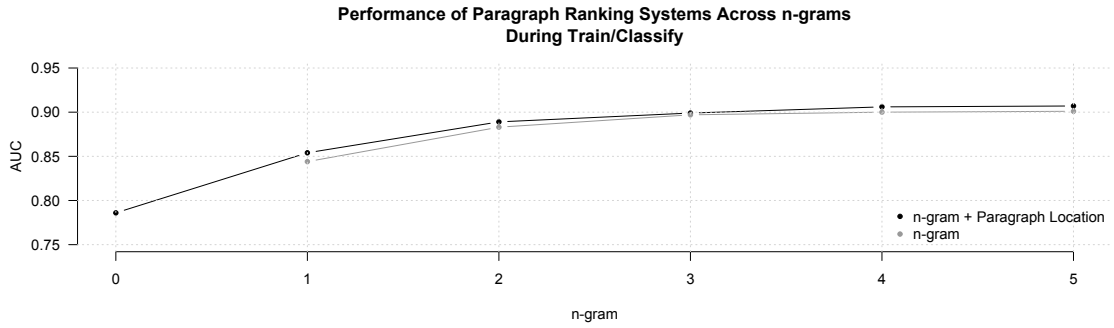

Figure 2: Performance of two configurations of the Finna system across increasing n-grams after training on the entire training data set, and classifying the hold-out testing set. The n-gram with paragraph location system was better than n-grams alone, but there was not a great difference.

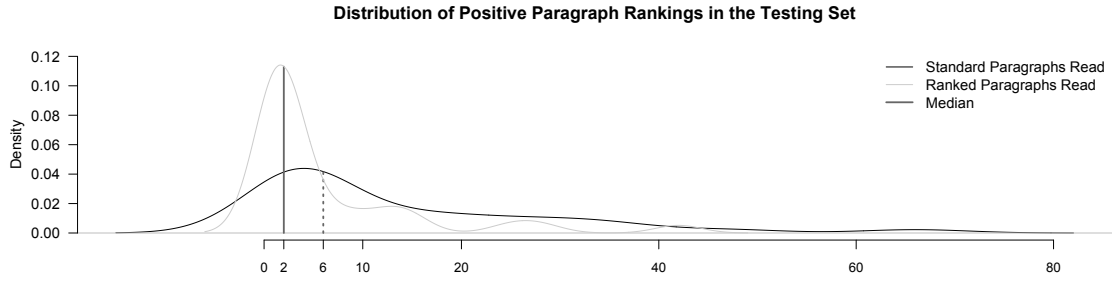

Figure 3: Distribution of positive-class paragraph rankings in the hold-out testing data set, after ranking based on a model trained on the training collection. The distribution of the number of paragraphs an annotator would have to read using the standard approach is shown in black, while the number of paragraphs he or she would have to read is shown in light grey. The median number for each distribution is shown as a thick dark grey line (dotted for the standard approach, and solid for the Finna system).

and Melissa Haendel, for contributing to the discussions that led to the experiments presented in this manuscript. Additionally, we wish to thank Giorgio Ascoli and David Hamilton, of the Neuron Registry, for their help in developing our data set.

References

Ambert, K., and Cohen, A. 2009. A system for classifying disease comorbidity status from medical discharge summaries using automated hotspot and negated concept detection. *Journal of the American Medical Informatics Association* 16(4):590.

Ambert, K., and Cohen, A. 2012. Text-mining and neuroscience. *International review of neurobiology* 103:109–132.

Ambert, K. 2013. *Text-mining Tools for Optimizing Community Database Curation Workflows in Neuroscience*. Ph.D. Dissertation, Oregon Health and Science University.

Ananiadou, S.; Rea, B.; Okazaki, N.; Procter, R.; and Thomas, J. 2009. Supporting systematic reviews using text mining. *Social Science Computer Review* 27(4):509–523.

Bandrowski, A. 2011. Biological resource catalog: Nif and neurolex.

Burge, S.; Attwood, T.; Bateman, A.; Berardini, T.; Cherry, M.; O'Donovan, C.; et al. 2012. *Biocurators and Biocuration: surveying the 21st century challenges*, volume 2012. Oxford University Press.

Burns, G.; Krallinger, M.; Cohen, K.; Wu, C.; and Hirschman, L. 2009. Studying biocuration workflows.

Cohen, A.; Ambert, K.; and McDonagh, M. 2009. Cross-topic learning for work prioritization in systematic review creation and update. *Journal of the American Medical Informatics Association* 16(5):690–704.

Cohen, A.; Ambert, K.; and McDonagh, M. 2010. A prospective evaluation of an automated classification system to support evidence-based medicine and systematic review. In *AMIA Annual Symposium Proceedings*, volume 2010, 121. American Medical Informatics Association.

Cohen, A.; Adams, C.; Davis, J.; Yu, C.; Yu, P.; Meng, W.; Duggan, L.; McDonagh, M.; and Smalheiser, N. 2010. Evidence-based medicine, the essential role of systematic reviews, and the need for automated text mining tools. In *Proceedings of the 1st ACM International Health Informatics Symposium*, 376–380. ACM.

Cohen, A. 2008. Five-way smoking status classification using text hot-spot identification and error-correcting output codes. *Journal of the American Medical Informatics Association* 15(1):32–35.

Fayyad, U., and Irani, K. 1993. Multi-interval discretization of continuous-valued attributes for classification learning.

Hirschman, L.; Burns, G.; Krallinger, M.; Arighi, C.; Cohen, K.; Valencia, A.; Wu, C.; Chatr-Aryamontri, A.; Dowell, K.; Huala, E.; et al. 2012. Text mining for the biocuration workflow. *Database: The Journal of Biological Databases and Curation*.

Joachims, T. 1998. Text categorization with support vector machines: Learning with many relevant features. *Machine learning: ECML-98* 137–142.

Karamanis, N.; Lewin, I.; Seal, R.; Drysdale, R.; and Briscoe, E. 2007. Integrating natural language processing with flybase curation. In *Pac Symp Biocomput*, volume 12, 245–56.

Pokkunuri, S.; Ramakrishnan, C.; Riloff, E.; Hovy, E.; and Burns, G. 2011. The role of information extraction in the design of a document triage application for biocuration. In *Proceedings of BioNLP 2011 Workshop*, 46–55. Association for Computational Linguistics.

Ramakrishnan, C.; Baumgartner Jr, W.; Blake, J.; Burns, G.; Cohen, K.; Drabkin, H.; Eppig, J.; Hovy, E.; Hsu, C.; Hunter, L.; et al. Building the Scientific Knowledge Mine (SciKnowMine1): a community-driven framework for text mining tools in direct service to biocuration. *New Challenges For NLP Frameworks Programme* 9.

Rodriguez-Esteban, R.; Iossifov, I.; and Rzhetsky, A. 2006. Imitating manual curation of text-mined facts in biomedicine. *PLoS Comput Biol* 2(9):e118.

Wallace, B.; Trikalinos, T.; Lau, J.; Brodley, C.; and Schmid, C. 2010. Semi-automated screening of biomedical citations for systematic reviews. *BMC bioinformatics* 11(1):55.

Wiegers, T.; Davis, A.; Cohen, K.; Hirschman, L.; and Mattingly, C. 2009. Text mining and manual curation of chemical-gene-disease networks for the Comparative Toxicogenomics Database(CTD). *BMC bioinformatics* 10(1):326.

Wiegers, T. 2009. Developing a Text Mining Prototype for the Comparative Toxicogenomics Database Biocuration Process.

Yang, J.; Cohen, A.; and McDonagh, M. 2008. SYRIAC: The SYstematic Review Information Automated Collection System A Data Warehouse for Facilitating Automated Biomedical Text Classification. In *AMIA Annual Symposium Proceedings*, volume 2008, 825. American Medical Informatics Association.