Automating Biomedical Evidence Synthesis: Recent Work and Directions Forward*

Byron C. Wallace[1]

College of Computer and Information Science
Northeastern University, Boston, MA
b.wallace@northeastern.edu

Abstract. Evidence-based medicine (EBM) looks to inform patient care with the totality of the available evidence. Systematic reviews, which statistically synthesize the entirety of the biomedical literature pertaining to a specific clinical question, are the cornerstone of EBM. These reviews are critical to modern healthcare, informing everything from national health policy to bedside decision-making. But conducting systematic reviews is extremely laborious, and hence expensive. Producing a single review requires thousands of expert hours, spent culling relevant structured evidence from the vast unstructured evidence base (i.e., natural language articles describing the conduct and results of trials). The exponential expansion of the biomedical literature base has exacerbated the situation: Health care practitioners can no longer keep up with the primary literature, and this hinders the practice of evidence-based care. The machine learning, natural language and information retrieval communities can lead the way in addressing this problem through the development of automation technologies that facilitate search and synthesis of evidence; but developing these will require meeting challenging technical problems. In this extended abstract, I discuss some of the progress made in recent years toward expediting unstructured biomedical evidence synthesis via automation techniques, and I highlight a few key challenges that remain.

Keywords: evidence-based medicine, natural language processing

How do we know which treatments actually work for particular patient populations, with respect to specified outcomes? Ideally, such decisions would be made on the basis of casual (relative) effect estimates of (comparative) treatment efficacies for outcomes of interest, typically derived from randomized controlled trials (RCTs) [10]. Regrettably, the results of such trials are typically disseminated via unstructured, natural language articles that describe findings. This makes it difficult to put evidence into practice.

* This abstract accompanies a keynote that touches upon a few threads of work, components of which are supported by the National Institutes of Health (grants R01LM012086 and UH2CA203711) and the National Science Foundation (CAREER 1750978). This is collaborative work with colleagues, including (inexhaustively) Iain Marshall, Ani Nenkova, Thomas Trikalinos and Matt Lease; as well as PhD students Ben Nye, Sarthak Jain, Roma Patel, Gaurav Singh, Ye Zhang and An Than Nguyen.

Researchers in machine learning (ML), natural language processing (NLP) and information retrieval (IR) can play a key role in making unstructured evidence more actionable, e.g., by facilitating search, extraction and ultimately synthesis of findings reported in articles that describe the outcomes of randomized controlled trials. Such approaches have the potential to afford healthcare providers access to the "best currently available evidence at the push of a button" [11]. Considerable progress has been made progress toward this aim [1] (not limited to work I have been involved with, of course, although this is what I focus on here). For example, colleagues and I have developed *RobotReviewer* [2], a prototype that integrates machine learning technologies to produce automated syntheses of the trials described in uploaded articles. However, despite this progress, core technical challenges remain. In this abstract, as in the talk it accompanies, I highlight some recent progress and select challenges that remain.

Training models in low-supervision settings. Inducing models that can automatically categorize and extract data from unstructured articles requires, of course, supervision on various fields of interests. State-of-the-art NLP models for relevant tasks such as information extraction tend to be highly parameterized neural networks and hence data hungry. It is difficult and expensive to acquire large volumes of training data in the biomedical domain: domain experts are few, busy and expensive, and articles describing clinical trials tend to be dense in jargon and hence difficult for lay annotators.

To address this challenge, we have explored a few avenues. The first is a paradigm of *distant supervision* [4], wherein 'found' data is re-purposed, typically via rules and heuristics, to provide noisy supervision for a target task. In particular we have exploited the Cochrane Database of Systematic Reviews (CDSR), a database of semi-structured data pertaining to individual articles, to derive such noisy supervision over sentences [3]. To mitigate noise, we have introduced an approach we call *Supervised Distant Supervision* [13] which harnesses a small amount of direct supervision to improve the quality of distantly derived labels. This improved the performance of a distantly supervised model for extracting clinically salient sentences in full-text articles [13].

Semi-supervised methods constitute a complementary approach to improving model performance in low-supervision settings. For instance, we were able to exploit structured abstracts to derive syntactic patterns that can be fed as additional inputs to sequence tagging models (e.g., LSTM-CRF) to yield improved performance [9]. And elsewhere, we have shown how to exploit existing ontologies/controlled vocabularies (e.g., MeSH) to impose inductive biases in neural models, in turn improving predictive accuracy [15].

Hybrid expert & crowd annotation. Another means of addressing a paucity of training data, of course, is to simply collect more data. As mentioned above, relying on biomedical domain experts for this would be prohibitively costly. And it is not obvious that layworkers (hired via crowdwork platforms like Amazon Mechanical Turk) will be able to perform the task. However, we have shown that redundant collection of annotations coupled with careful aggregation strategies yields reasonable training signal [6, 5]. And we have recently made publicly

available a relatively large set (∼5k) of richly annotated biomedical abstracts of papers describing clinical trials to facilitate methodological work on NLP for EBM [8].

It is not obvious how best to jointly exploit small amounts of (pricey) expert supervision and (cheap but noisy) crowd annotations at scale. We have explored active approaches for this [7], but believe the general problem remains ripe for exploration, especially in regards to also incorporating machine predictions in the loop [14].

Joint extraction & inference over lengthy documents. Ideally, we would like to cull from article full-texts assertions that the underlying trial described in a given article provides evidence in favor of a particular treatment for a specified condition and outcome. This requires jointly extracting these fields and then inferring what has been reported regarding them. In general, extracting relationships between entities in scientific papers remains an exciting open challenge at the fore of existing language technologies [12].

Concerning the particular domain of EBM, we are just beginning work on assembling a corpus that will comprise pairs of 'evidence frames' specifying an intervention, a comparator, and an outcome and accompanying full-text articles. The task, then, will be to predict whether the article provides evidence that the given intervention is more effective than the comparator, with respect to the outcome (or not). Going forward, we envision a model that can simultaneously extract the interventions, comparators and outcomes studied (e.g., trained using the corpus mentioned above [8]) *and* infer the reported directionality of the findings. This is an audacious goal, but if realized would afford access to immediately actionable evidence, automatically.

Closing remarks. The above are just a sample of the challenges inherent to the task of trying to trying to automate biomedical evidence synthesis. In addition to discussing work I have done with colleagues toward meeting these, my aim in this talk and extended abstract is to call attention to the general problem of evidence synthesis; I think researchers in IR and adjacent areas have the potential to change the practice of evidence-based medicine by helping doctors navigate the evidence, and ultimately figure out what works. This is a nice general problem to work on because it is both socially important and technically challenging.

References

1. Jonnalagadda, S.R., Goyal, P., Huffman, M.D.: Automating data extraction in systematic reviews: a systematic review. Systematic reviews **4**(1), 78 (2015)
2. Marshall, I.J., Kuiper, J., Banner, E., Wallace, B.C.: Automating biomedical evidence synthesis: Robotreviewer. In: Proceedings of the conference. Association for Computational Linguistics. Meeting. vol. 2017, p. 7. NIH Public Access (2017)
3. Marshall, I.J., Kuiper, J., Wallace, B.C.: Automating risk of bias assessment for clinical trials. Biomedical and Health Informatics, IEEE Journal of **19**(4), 1406–1412 (2015)
4. Mintz, M., Bills, S., Snow, R., Jurafsky, D.: Distant supervision for relation extraction without labeled data. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2. pp. 1003–1011. Association for Computational Linguistics (2009)
5. Mortensen, M.L., Adam, G.P., Trikalinos, T.A., Kraska, T., Wallace, B.C.: An exploration of crowdsourcing citation screening for systematic reviews. Research synthesis methods **8**(3), 366–386 (2017)
6. Nguyen, A.T., Wallace, B.C., Li, J.J., Nenkova, A., Lease, M.: Aggregating and predicting sequence labels from crowd annotations. In: Proceedings of the conference. Association for Computational Linguistics. Meeting. vol. 2017, p. 299. NIH Public Access (2017)
7. Nguyen, A.T., Wallace, B.C., Lease, M.: Combining crowd and expert labels using decision theoretic active learning. In: Third AAAI Conference on Human Computation and Crowdsourcing (2015)
8. Nye, B., Li, J.J., Patel, R., Yang, Y., Marshall, I.J., Nenkova, A., Wallace, B.C.: A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. Association for Computational Linguistics (ACL) (2018)
9. Patel, R., Yang, Y., Marshall, I., Nenkova, A., Wallace, B.: Syntactic patterns improve information extraction for medical search. arXiv preprint arXiv:1805.00097 (2018)
10. Sackett, D.L.: Evidence-based Medicine How to practice and teach EBM. WB Saunders Company (1997)
11. Tsafnat, G., Dunn, A., Glasziou, P., Coiera, E., et al.: The automation of systematic reviews. BMJ **346**(f139), 1–2 (2013)
12. Verga, P., Strubell, E., Shai, O., McCallum, A.: Attending to all mention pairs for full abstract biological relation extraction. arXiv preprint arXiv:1710.08312 (2017)
13. Wallace, B.C., Kuiper, J., Sharma, A., Zhu, M.B., Marshall, I.J.: Extracting pico sentences from clinical trial reports using supervised distant supervision. Journal of Machine Learning Research **17**(132), 1–25 (2016)
14. Wallace, B.C., Noel-Storr, A., Marshall, I.J., Cohen, A.M., Smalheiser, N.R., Thomas, J.: Identifying reports of randomized controlled trials (rcts) via a hybrid machine learning and crowdsourcing approach. Journal of the American Medical Informatics Association **24**(6), 1165–1168 (2017)
15. Zhang, Y., Lease, M., Wallace, B.C.: Exploiting domain knowledge via grouped weight sharing with application to text categorization. arXiv preprint arXiv:1702.02535 (2017)