Book Review

Statistical Significance Testing for Natural Language Processing

Rotem Dror, Lotem Peled-Cohen, Segev Shlomov, and Roi Reichart
(Technion Israel Institute of Technology)

Morgan & Claypool (Synthesis Lectures on Human Language Technologies, edited by Graeme Hirst, volume 45), 2020, xx+98 pp; paperback, ISBN 978-1-68173-795-9, $49.95; ebook, ISBN 978-1-68173-796-6, $39.96; hardcover, $69.95;
doi:10.2200/S00994ED1V01Y202002HLT045

Reviewed by
Edwin D. Simpson
University of Bristol

Like any other science, research in natural language processing (NLP) depends on the ability to draw correct conclusions from experiments. A key tool for this is statistical significance testing: We use it to judge whether a result provides meaningful, generalizable findings or should be taken with a pinch of salt. When comparing new methods against others, performance metrics often differ by only small amounts, so researchers turn to significance tests to show that improved models are genuinely better. Unfortunately, this reasoning often fails because we choose inappropriate significance tests or carry them out incorrectly, making their outcomes meaningless. Or, the test we use may fail to indicate a significant result when a more appropriate test would find one. NLP researchers must avoid these pitfalls to ensure that their evaluations are sound and ultimately avoid wasting time and money through incorrect conclusions.

This book guides NLP researchers through the whole process of significance testing, making it easy to select the right kind of test by matching canonical NLP tasks to specific significance testing procedures. As well as being a handbook for researchers, the book provides theoretical background on significance testing, includes new methods that solve problems with significance tests in the world of deep learning and multidataset benchmarks, and describes the open research problems of significance testing for NLP.

The book focuses on the task of comparing one algorithm with another. At the core of this is the **p-value**, the probability that a difference at least as extreme as the one we observed could occur by chance. If the p-value falls below a predetermined threshold, the result is declared significant. Leaving aside the fundamental limitation of turning the validity of results into a binary question with an arbitrary threshold, to be a valid statistical significance test, the p-value must be computed in the right way. The book describes the two crucial properties of an appropriate significance test: The test must be both *valid* and *powerful*. Validity refers to the avoidance of *type 1* errors, in which the result is incorrectly declared significant. Common mistakes that lead to type 1 errors include deploying tests that make incorrect assumptions, such as independence between data points. The power of a test refers to its ability to detect a significant result and therefore to avoid *type 2* errors. Here, knowledge of the data and experiment must be used to choose a test that makes the correct assumptions. There is a trade-off between validity and power, but for the most common NLP tasks (language modeling, sequence labeling, translation, etc.), there are clear choices of tests that provide a good balance.

https://doi.org/10.1162/coli_r_00388

Beginning with a detailed background on significance testing, the book then shows the reader how to carry out tests for specific NLP tasks. There is a mix of styles, with the first four chapters providing reference material that will be extremely useful to both new and experienced researchers. Here, it is easy to find the material related to a given NLP task. The next two chapters discuss more recent research into the application of significance tests to deep neural networks and for testing across multiple datasets. Alongside open research questions, these later chapters provide clear guidelines on how to apply the proposed methods. It is this mix of background material and reference guidelines that I believe makes this book so compelling and nicely self-contained.

The introduction in Chapter 1 motivates the need for a comprehensive textbook and outlines challenges that the later chapters address more deeply. The theoretical background material begins in Chapter 2, which introduces core concepts, including hypothesis testing, type 1 and type 2 errors, validity and power, and p-values. The reader does not need to have any prior knowledge of statistical significance tests to follow this part. However, experienced readers could still benefit from reading this chapter, as concepts such as p-values are widely misunderstood and misused (Amrhein, Greenland, and McShane 2019).

The significance tests themselves are introduced in Chapter 3, categorized into parametric and nonparametric tests. The chapter begins with the intuitively simple *paired z-test*, then builds up to more commonly-applied techniques, showing the connections and assumptions that each test makes. Step-by-step algorithms help the reader to implement each test. Although the chapter does cite uses of tests in NLP research, the main purpose is to present the theory behind each test and point out their differences.

Chapter 4 provides perhaps the most handy part of the book for reference: a correspondence between common NLP tasks and statistical tests. Each task is discussed in terms of the evaluation metrics used, then a decision tree is introduced to guide the reader toward a choice between a parametric test, bootstrap or randomization test, or sampling-free nonparametric test. Section 4.3 then links each NLP evaluation measure to a specific significance test, presenting a large table that helps readers identify which test they need for a specific task. Particular considerations for each task are also pointed out to provide more detail about the appropriate options. The final part of this chapter describes the issue of **p-hacking**, in which dataset sizes are increased until a significance threshold is reached without consideration for biases in the data (discussed, for example, in Hofmann [2015]). The chapter proposes a simple solution to ensure robust significance testing with large datasets.

Where Chapter 4 presents well-established methods, Chapter 5 introduces the current research question of how best to apply statistical significance testing to deep learning. Non-convex loss functions, stochastic optimization, random initialization, and a multitude of hyperparameters limit the conclusions we can draw from a single test run of a deep neural network (DNN). This chapter, which is based on the authors' ACL paper (Dror, Shlomov, and Reichart 2019), explains how the comparison process can be overhauled to provide more meaningful evaluations. Beginning by explaining the difficulties of evaluating DNNs, the chapter then introduces criteria for a comparison framework, then discusses the limitations of current methods. Reimers and Gurevych (2018) have previously tackled this problem, but their approach has limited power and does not provide a confidence score. Other works, such as Clark et al. (2011), compare DNNs using a collection of statistics, such as the mean or standard deviation of performance across runs. This book shows how such an approach violates the assumptions of the significance tests. The authors propose *almost stochastic dominance* as the basis for a

better alternative. The chapter explains how to use the proposed method, evaluates it in an empirical case study, and finally analyzes the errors made by each testing approach.

Large NLP models are often tested across a range of datasets, which presents another problem for standard significance testing. Chapter 6 discusses the challenges of assessing two questions: (1) On how many datasets does algorithm A outperform algorithm B? (2) On which datasets does A outperform B? Applying standard significance tests individually to each dataset and counting the number of significant results is likely to overestimate the total number of significant results, as this chapter explains. The authors then present a new framework for replicability analysis, based on *partial conjunction testing*, and discuss two variants (Bonferroni and Fisher) for when the datasets are independent or dependent. They introduce a method based on Benjamini and Heller (2008) to count the number of datasets where one method outperforms another, then show how to use the Holm procedure (Holm 1979) to identify which datasets these are. Chapter 6 provides a lot of detailed background on the proposed replicability analysis framework, and the later sections again link the process to specific NLP case studies, and step-by-step summaries help the reader to apply the methodology. Extensive empirical results illustrate the very substantial differences in outcomes between the proposed approach and standard procedures.

The final two chapters present open problems and conclude, showing that the topic has many interesting research questions of its own, such as problems when performing cross-validation, and the limited statistical power of replicability analysis.

Overall, I highly recommend this book to a wide range of NLP researchers, from new students to seasoned experts who wish to ensure that they compare methods effectively. The book is excellent as both an introduction to the topic of significance testing and as a reference to use when evaluating your results. For anyone with further interest in the topic, it also points the way to future work. If one could level any criticism at this book at all, it is that it does not deeply discuss the basic flaws of significance testing or what the alternatives might be. For now, though, significance testing is an integral part of NLP research and this book provides a great resource for researchers who wish to perform it correctly and painlessly.

References

Amrhein, Valentin, Sander Greenland, and Blake McShane. 2019. Scientists rise up against statistical significance. Mar; 567(7748):305–307. DOI: https://doi.org/10.1038/d41586-019-00857-9, PMID: 30894741

Benjamini, Yoav and Ruth Heller. 2008. Screening for partial conjunction hypotheses. *Biometrics*, 64(4):1215–1222. DOI: https://doi.org/10.1111/j.1541-0420.2007.00984.x, PMID: 18261164

Clark, Jonathan H., Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 176–181, Portland, OR.

Dror, Rotem, Segev Shlomov, and Roi Reichart. 2019. Deep dominance—how to properly compare deep neural models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2773–2785, Florence. DOI: https://doi.org/10.18653/v1/P19-1266

Hofmann, Marko A. 2015. Searching for effects in big data: Why p-values are not advised and what to use instead. In *2015 Winter Simulation Conference (WSC)*, pages 725–736, IEEE.

Holm, Sture. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70. DOI: https://doi.org/10.1109/WSC.2015.7408210, PMID: 24482542

Reimers, Nils and Iryna Gurevych. 2018. Why comparing single performance scores does not allow to draw conclusions about machine learning approaches. *arXiv preprint arXiv:1803.09578*.

Edwin D. Simpson is a Lecturer in the Department of Computer Science, University of Bristol, UK. His research focuses on natural language processing (NLP), with particular interest in applying interactive machine learning and Bayesian techniques to NLP problems such as argumentation, crowdsourced annotation, and text ranking. His e-mail address is edwin.simpson@bristol.ac.uk.