Vol.2016-NL-229 No.18
2016/12/22

評価実験の設計と論文での結果報告: きちんとやっていますか?

酒井 哲也[1,a)]

概要：本講演では，まず古典的統計学に基づく検定手法，特に t 検定と分散分析について復習し，次に統計的検定に関する誤解や限界，および実験結果の適切な報告の仕方について論じる．さらに，情報検索・情報アクセス研究における事例を示しながら，研究者がシステム評価のために新たなテストコレクションを構築する際，および従来研究の結果をもとに新たな実験を行う際の一般的なサンプルサイズ設計方法を紹介する．自然言語処理研究におけるサンプルに基づく評価においても，これらの手法の導入により適切な検出力をもつ実験が適切な形で報告されるようになれば，個々の研究成果が有機的につながり真の知見が蓄積されるようになると考える．

キーワード：効果量, 評価, サンプルサイズ, 検出力, 統計的検定

Designing Experiments and Reporting Results: Are You Doing It Correctly?

Tetsuya Sakai[1,a)]

Abstract: In this talk, I will first review classical statistical significance tests, t tests and Analysis of Variance (ANOVA) tests in particular, and then discuss common misunderstandings and limitations of significance tests, as well as how researchers should report on experimental results. Moreover, while providing actual examples from information retrieval and access research, I will describe known methods for determining the sample size for constructing a new test collection for system evaluation, and for conducting a new experiment based on results from previous work. I argue that, if sample-based evaluations in natural language processing research take up these practices in order to conduct experiments with appropriate levels of statistical power and to report results appropriately, individual research outomes could be integrated effectively for accumulating true knowledge.

Keywords: effect sizes, evaluation, sample sizes, statistical power, statistical significance.

[1] 早稲田大学
Waseda University
[a)] tetsuyasakai@acm.org