Practical Significance Testing for Experiments in Natural Language Processing

Um zwei unterschiedliche Systeme korrekt zu vergleichen, muss man die Daten in training, dev und test Data aufteilen. Wenn bei einem Signifikanztest das eine System bessere Ergebnisse als das andere liefert, stellt sich die Frage, ob die Unterschiede zufällig sind oder das neue System wirklich besser ist. Oft wird nur gefragt, ob die Systeme verschieden sind. Dies kann man mit Hilfe der Null Hypothese bestätigen bzw. widerlegen. Die Differenz ist signifikant gdw. p-value $\leq \alpha$. Typische Werte für Signifikanzniveau α sind: 0.05, 0.01, 0.001 . . .
Man kann verschiedene Metriken mit unterschiedlichen Tests ausdrucken: Accuracy (sign test) Mean Average Precision (Paired t-test), F-Score (randomized tests), Testen von Annahmen über Data (Chi-square test) usw.