Zusammenfassung Anwendung von Signifikanztests für Experimente in NLP

Bei einem Signifikanztest werden zwei verschiedene Systeme miteinander verglichen. Ist ein System besser als das Andere, bzw. sind Systeme verschieden? Jedes System hat eine Ausgabe. Die Ausgaben werden mithilfe von Evaluationsmaßen verglichen. Dann misst man den Unterschied zwischen den beiden Ausgaben. Ist der Unterschied signifikant, kann man die Nullhypothese verwerfen, ansonsten nicht. Maße für die Evaluierung sind Accuracy, z.B. binomial Test, Mean Average Precision (MAP), z.B. t-test, F-Score, z.B. randomized Tests und Tests von Annahmen über Daten, z.B. Chi-Quadrat Test. Bei Accuracy wird überprüft ob sich die Anzahl der korrekten Ausgaben signifikant verändert hat. MAP vergleicht eine Menge von Listen die von zwei Systemen eingestuft wurden und überprüft ob die durchschnittliche Einstufung eine signifikante Veränderung aufweist. Die F-Score ist schwierig zu interpretieren, da die Gesamtbewertung bei einzelnen Tests nicht linear ist. Tests von Annahmen über Daten können für das Finden von Features benutzt werden. Um Fallstricke bei Signifikanztests zu vermeiden, sollte man das Ergebnis korrekt und nicht einseitig wiedergeben. Fallstricke entstehen bei häufigen Tests.