

Es wurde über praktische Testing für Experimente besprochen. Alle Daten sind auf Training Set, Development Set und Test Set geteilt. Mit Training Set kann man die Parameter automatisch zu erlernen, mit Test Set kann man verschiedene Sachen ausprobieren (z.B. Klassifikators, Toolkits um zu schauen was am besten funktioniert), mit Test Set definiert man den Schätzwert für Performance.

Damit kann man zwei verschiedene Systeme vergleichen. Statistische Significance Tests soll man durchzuführen um herauszufinden ob die unterschiedliche Resultate liegen daran daß ein System ist besser als andere oder der Unterschied durch den Zufall passiert.

Für Testing verschiedene Metrics braucht man z.B. Accuracy, Mean Average Precision (Ranking-Maßnahme häufig in Retrieval-Evaluierung verwendet). Für Testing der Accuracy zählt man Trials und Successes - dass sind die Anzahl der Fälle, in denen A und B sich unterscheiden und entsprechend die Anzahl der Fälle wo B ist besser als A.