Klyueva Olga
CIS, LMU München
29.05.1917

Computerlinguistisches Arbeiten: Protokoll zum Repetitorium der Computerlinguistik.
Practical significance testing for experiments in natural language processing.

Als erstes wurde über Developmentset, Trainingset und Testset gesprochen. An einem Developmentset probiert man verschiedene Klassifikatoren aus und stellt fest, was am besten funktioniert. Ein Trainingset ist dafür da, die Parameter automatisch zu lernen. Man wählt beispielweise die Klassifikatoren, die man am Ende vergleichen will, und probiert sie einmal auf Testdaten.
Als nächstes wurde ein Beispiel gegeben. Die Daten wurden richtig in Trainings-, Development- und Testdaten aufgeteilt. Es gibt zwei Systeme A und B, die die Ausgaben an die Testdaten liefern. Evaluation Metrik zeigt, dass System B einen höheren Score als System A hat. Die Frage lautet, ob System B besser als System A ist und ob System B auf andere Daten bessere Resultate liefert als System A. Wenn die Wahrscheinlichkeit, dass die beobachteten Unterschiede zufällig entstanden sind, geringer als Alpha ist, dann ist der Unterschied zwischen zwei Systemen signifikant groß. Wenn der Wahrscheinlichkeitswert höher als Alpha ist, dann ist das System nicht besser als System A. In diesem Fall ist der Unterschied zwischen zwei Systemen zufällig oder gibt es nicht genug Daten.
Als nächstes wurde noch ein Beispiel mit zwei Systemen A und B, die beobachteten Ergebnisse und Metriken präsentiert. Die Frage ist, ob es einen Unterschied zwischen den zwei Systemen gibt. Null Hypothese ist, dass es keinen Unterschied zwischen zwei Systemen von der erwartbaren Performance gibt. Test Statistik ist ein Maß, das zeigt, dass es einen Unterschied zwischen beiden Evaluation-Metriken gibt. Es wurde eine Verteilung über die Unterschiede zwischen zwei Systemen gefunden unter der Annahme, dass die Null Hypothese wahr ist.

Als nächstes wurde über eine Metrik Accuracy gesprochen. Es wurde ein Beispiel gezeigt mit zwei Systemen: System A und System B. Es gibt 10 Fälle, wo das System entweder richtige oder falsche Resultate erzeigt hat. Wenn es das System richtig gemacht hat, wird es als 1 kodiert, wenn falsch als 0. Das System A hat die meisten Fälle falsch und zwei Fälle richtig. Das System B hat in zwei Fällen falsch. System A hat eine Accuracy von 20 Prozent, System B eine Accuracy von 80 Prozent.
Für den Sign-Test braucht man zwei Parameter:

- In wie vielen Fällen unterscheidet sich System A von System B: 8
- In wie vielen Fällen ist System B besser als System A: 7

Mit dem Python Befehl bekommt man ein Resultat von 7 Prozent. Es ist grösser als Alpha mit 5 Prozent. Daraus folgt, dass das System B nicht signifikant besser als System A ist. Trotz der Tatsache, dass der Unterschied in den Accuracy-Werten der beiden Systeme groß ist, ist das System B nicht signifikant besser als System A. Das ist deswegen, weil die Datenlage zu klein ist. 10 Fälle reichen nicht, um eine Aussage zu treffen. Man braucht mindestens 50 Fälle.
Die nächste Metrik ist mean average precision. Diese Metrik benutzt man in Information Retrieval Evaluierung. Der User gibt ein Query zur Suchmaschine. Die Suchmaschine gibt relevante Dokumente aus. Mean average precision berechnet, wie gut das Ranking von Dokumenten ist. Es wurde ein Beispiel mit zwei Systemen präsentiert: System A und System B. Es gibt 10 Queries mit einem Ranking-Wert per Query. Mean avarage precision für System A ist 30 Prozent, für System B 65 Prozent. Es wurden zwei Listen mit den Werten für System A und B gemacht. Mit dem Python Befehl rechnet man den p-Wert aus. Der P-Wert in diesem Fall ist gleich 3,78 Prozent. Er ist kleiner als Alpha. Die Resultate sind signifikant.

Als nächstes wurde eine Zusammenfassung für paired T-Test und Sign Test erwähnt.
Paired T-Test:

- Compare scores für matching pairs.
- Die Annahme ist, dass der Unterschied zwischen den Systeme Gaussian verteilt ist
- Die Null Hypothese ist, dass der Mittelwert der Gaussian Verteilung null ist.

Sign Test:

- Testfälle werden als Binärentscheidungen angeschaut. Es wird die Wahrscheinlichkeit betrachtet, mit der sich die Testfälle mit der neuen Methode verbessern.

- Die Null Hypothese ist, dass die Verbesserungen zufällig sind mit einer Wahrscheinlichkeit von 50 Prozent.

Mit beiden Methoden werden Duplikaten von Testdaten entfernt.

Am Schluss wird pitfalls of significant testing anhand Comic präsentiert. Die Wissenschaftler sagen, dass Akne und Jelly Bins positiv korreliert sind, aber der p-Wert wird grösser als 5 Prozent. Es ist nicht signifikant. Man macht immer mehr Experimente über verschiedene Farben. Die Beobachtungen sind immer nicht signifikant. Irgendwann hat man das signifikante Resultat. Eine Zusammenfassung lautet: je mehr Experimente man macht, desto mehr type 1 Fehler hat man, die Signifikanz andeuten, wo nichts Signifikantes zu beobachten ist. In diesem Fall soll man Bonferroni Correction machen, wo man den P-Wert durch die Anzahl der Experimente teilt, die man vorhat.