Protokoll zur Sitzung am 29.5.17 (Repetitorium)

Im Laufe des heutigen Repetitoriums wurde das Thema „Signifikanztests der Experimenten in NLP“ behandelt.

In NLP ist es oft nötig, mehrere Systeme miteinander zu vergleichen. Insbesondere geht es darum, ob ein System tatsächlich besser als ein anderes ist, wenn man sich nicht nur auf einen einzigen Test beschränkt. Oder beruht die bessere Performance eines Systems auf Zufall?

Um diese Frage zu beantworten, muss man einige Schritte durchführen, die im Rahmen eines Algorithmus (*Statistical Significance Tests Framework*) vorgestellt wurden:

- Sei System B in irgendeiner Metrik besser, als System A. Dann:

    1) Als *Null-Hypothese* gilt: Es gibt keinen Unterschied zwischen den beiden Systemen.
    2) Die Unterschiede zwischen den Ausgaben beider Systeme müssen *quantifiziert* werden.
    3) Man muss eine *Distribution* über diese Unterschiede finden (unter der Annahme, dass die Null-Hypothese gilt).
    4) Am Ende soll man schauen, ob:
        - es wahrscheinlich ist, dass es noch mehr Extreme geben kann ($\rightarrow$ der Unterschied ist **nicht** signifikant).
        - es eher unwahrscheinlich ist ($\rightarrow$ der Unterschied **ist signifikant**).

Was den 3. Schritt angeht, wurde hauptsächlich über 2 Verfahren berichtet: „paired t-test“ und „sign test“. Anhand eines Beispiels wurde außerdem deutlich gemacht, dass 2 Systeme, die sehr unterschiedliche Performance haben, eben *nicht* signifikant abweichende Resultate aufweisen können.

Am Ende des Repetitoriums wurde betont, dass man „Typ I Fehler“ (Signifikanz da andeuten, wo es nicht der Fall ist) vermeiden soll, indem man insbesondere ehrlich über den Verlauf seiner Forschung berichtet.