PRACTICAL SIGNI CANCE TESTING FOR EXPERIMENTS IN NATURAL LANGUAGE PROCESSING

Anton Serjogin
Centre for Information and Speech Processing, LMU
anton.serjogin@gmail.com

29.05.2017

When using different systems we can get different outputs and the metrics of those can differentiate. This leads us to the following question - either one of the system has better performance or this is due to chance. In order to determine this occurrence, statistical significance tests are needed. *Null hypothesis* is one way of observing the difference of the output data. Metrics, such as *precisiony* (number of correct answers given by the system / number of answers given by the system), *recall* (number of correct answers given by the system / total number of correct answers), *f-score* show us the performance levels of the system. A very important aspect is suggesting significance, because the more tests we run, the more errors we may get, when there is no significance declared.