11 Natural Language Processing (ES)

The distributional hypothesis states that the meaning of a word can be defined by its use and, therefore, it can be represented as a distribution of contexts in which the word occurs in a large text corpus.

(*a*) Describe four different types of context that can be used for this purpose. [4 marks]

(*b*) The contexts can be weighted using Pointwise Mutual Information (PMI). Explain, giving formulae, how PMI is calculated and how individual probabilities are estimated from a text corpus. [5 marks]

(*c*) Some words occur very rarely in the corpus. How does this affect their PMI scores as contexts? [4 marks]

(*d*) The goal of distributional word clustering is to obtain clusters of words with similar or related meanings. The following clusters have been produced in two different noun clustering experiments:

Experiment 1:

carriage bike vehicle train truck lorry coach taxi
official officer inspector journalist detective constable policeman reporter
sister daughter parent relative lover cousin friend wife mother husband brother father

Experiment 2:

car engine petrol road driver wheel trip steering seat highway sign speed
concert singer stage light music show audience performance ticket
experiment research scientist paper result publication laboratory finding

(*i*) How are the clusters produced in the two experiments different with respect to the similarity they capture? What lexico-semantic relations do the clusters exhibit? [3 marks]

(*ii*) The same clustering algorithm, K-means, was used in both experiments. What was different in the setup of the two experiments that resulted in the different kinds of similarity captured by the clusters? [4 marks]