Guide for Reproducibility of the LocalMaxs Improvement and its Results

February 19, 2023

(*Unidentified authors at the paper submission stage*)

1 Introduction

This is a guide to reproduce the algorithms related to the content of the paper titled *Improving LocalMaxs Multiword Expression Extractor*. File **Archive.zip** must be uncompressed. It contains the python files

LocalMaxs.py,
StopWordsModule.py.

And the following files, which are text *corpora*:

EN6.0Mw,
EN0.5Mw,
FR6.1Mw,
PT6.1Mw,
DE6.0Mw.

In order to perform reproducibility, all these files may be dowloaded to a single folder.

2 Extracting Relevant Expressions from *Corpora*

In order to extract Relevant Expressions (RE) from *corpora*, the python function **LocalLocalMaxs(CorpusFileName, MaxNgramLength, AllMetrics, ResSampleForEval, P)** must be called. These are the meaning of the parameters:

CorpusFileName is the name of the unziped text file corresponding to the text *corpus* from where REs will be extracted;

MaxNgramLength, whose default value equals 7, is the greatest number (plus one) of words forming the REs to be extracted from the *corpus*;

AllMetrics, which is a boolean parameter, indicates whether or not the user wants to obtain four different lists of REs extracted, each one based on one of the four cohesion functions: $\chi^2\_f(.)$; $SCP\_f(.)$; $Dice\_f(.)$ and $MI\_f(.)$. The default is False and means that only $\chi^2\_f(.)$ will be used;

ResSampleForEval, which is boolean, indicates whether or not the user wants to obtain random samples, each one from the full set of REs extracted by the LocalMaxs using one of the four cohesion functions. The size of each sample set is 752, in order to ensure a 95 % confidence level on a confidence interval equal to 0.7 ± 0.033, for a *z-score*$=1.96$ and a proportion value (Precision in this case) around 0.7. Default is False, producing no sample.

P is a positive integer for the power of the generalized mean used in the criterion for the selection of the REs, defined in improved LocalMaxs proposed in Definition 2 of the paper. Default value is 2.

So, this means that taking for example a file named "CorpusFile", if *LocalMaxs("CorpusFile")* is called in python context, only a file named "REsQuiSqr_CorpusFile_2" is produced containing the full set of REs, from 1-grams to 6-grams, extracted from the *corpus* in "CorpusFile" file, each line containing an RE. If the command is *LocalLocalMaxs("CorpusFile", P=1)*, a file named "REsQuiSqr_CorpusFile_1" will be generated with REs selected using $P = 1$ for the power of the referred generalised mean.

In order to estimate the Precision of the REs extracted from a *corpus* by the LocalMaxs using one of the four cohesion measures, a random sample of those REs must be evaluated for that. So, in addition to the Recall, if the user wants, for example, to evaluate the Precision of the REs extracted from the "EN6.0Mw" (one of available *corpora* as referred) according to each of the cohesion metrics and using P=2, *LocalLocalMaxs("EN6.0Mw", AllMetrics=True, ResSampleForEval=True)* must be called. In this case, the following files will be generated: "REsQuiSqr_EN6.0Mw_2", "REsQuiSqr_EN6.0Mw_2_Sample", "REsSCP_EN6.0Mw_2", "REsSCP_EN6.0Mw_2_Sample", "REsDice_EN6.0Mw_2", "REsDice_EN6.0Mw_2_Sample", "REsMI_EN6.0Mw_2" and "REsMI_EN6.0Mw_2_Sample".

— End of the Guide —