Documentation of Alignment and Report Capabilities

<Oct 11, 2021>

Summary

- Analyze for cluster and partial clustering
- Reports from resulting analysis file.
- Example reports for BEST conditions observed.
- Note that partial analysis option used cluster scorer for lexstat (due to error using partial scorer).
- Pairwise alignment and reports.
- Example reports for BEST conditions observed.

Accessing KeyPano commands

- Must have keypano installed which includes cldfbench, lexibank, …
  - Analysis and reporting commands also support running direct from the command.
- Cldfbench has list of commands for making cldf or updating…

Commands:

Cldfbench commands

- cldfbench keypano.<command> <command-args>
- download - downloads the starting database from which to compose the cldf
  - cldfbench download # runs download command using keypano.py
- makecldf - uses etc/orthograpy files to construct database with segments IPA.
  - cldfbench lexibank.makecldf # makes cldf database from IDS tables and orthography profiles.

Cldf analysis and report commands:

- pairwise donor report versus Spanish and Portuguese language donors.

```
 % cldfbench keypano.report_donor --model asjp --
threshold 0.275 --series "pairwise for release" >
output/pairwise-0.275-asjp.txt
```

- cluster analysis with sca scorer

```
% cldfbench keypano.analyze cluster --model sca --method
sca --mode global --cluster_method upgma --threshold
0.25 --series "cluster-for-release" > output/
cluster-0.25-sca-sca.txt
```

- cluster analysis with lexstat scorer

```
 % cldfbench keypano.analyze cluster --model sca --
method lexstat --runs 10000 --mode global --
cluster_method upgma --threshold 0.55 --series "cluster-
for-release-lex" > output/cluster-0.55-sca-lex.txt
```

- reports for sca scorer analysis
- % cldfbench keypano.report --store store --infile cluster-for-release-sca.tsv --exclude Indo-European --family Pano-Tacanan --status ntn --series "pano-sca-global-upgma" > output/cluster-0.25-pano-sca-summary.txt
- % cldfbench keypano.report --store store --infile cluster-for-release-sca.tsv --exclude Indo-European --status ntn --series "sca-global-upgma" > output/cluster-0.25-sca-summary.txt

- reports for lexstat scorer analysis
- % cldfbench keypano.report --store store --infile cluster-for-release-lex.tsv --exclude Indo-European --status ntn --series "lex-global-upgma" > output/cluster-0.55-lex-summary.txt
- % cldfbench keypano.report --store store --infile cluster-for-release-lex.tsv --exclude Indo-European --family Pano-Tacanan --status ntn --series "pano-lex-global-upgma" > output/cluster-0.55-pano-lex-summary.txt

- reports for pairwise
- % cldfbench keypano.report_donor --model asjp --threshold 0.275 --limit 0.75 --status ntn --series "rl0.75-asjp" > output/pairwise-0.275-rl0.75-asjp.txt

BEST Conditions based on detection performance and IDS annotation of borrowing

<Oct 5, 2021>

- Pairwise
  - Threshold - 0.275 (could round to 0.3)
  - Mode - no difference between overlap or global
  - Model - asjp is optimal, sca slightly poorer

    - Predicted F1 = 0.60, prec=0.70, recall = 0.50
- Cluster only, partial not competitive
    - sca model, sca method
        - mode ‘global’
        - cluster method upgma
        - threshold 0.25
        - F1 = 0.48, precision = 0.42, recall = 0.55
    - sca model, lexstat method
        - mode ‘global’
        - runs 10,000 for better precision
        - cluster method upgma
        - threshold 0.55
        - F1=0.50, precision = 0.5, recall = 0.5

Analyze

Based on original analysis module adapted by Mattis from Seabor. Uses cluster analysis to identify possible cognate groups and distinguishes between groups internal to a language family and groups that cross language families.

Arguments for analysis:

- Module: ‘cluster’ or ‘partial’ clustering. Default is ‘cluster’.
    - Partial is not viable in present analysis giving much poorer results than cluster.
- Method: ‘sca’ or ‘lexstat’ scoring. Default is ‘lexstat’.
- Model: ’sca’ or ’asjp’ sound class model. Default is ‘sca’.
- Threshold: threshold limits below which cognate clusters are identified. Default is 0.5. Higher is less restrictive and results in more false positives.
- Mode: ‘global’, ‘local’, ‘overlap’ or ‘dialign’ methods of alignment. Default is ‘overlap’.
- Cluster_method: ‘upgma’ or ‘infomap’ clustering method. Default is ‘infomap’. ‘Infomap’ is more compute intensive performing randomization tests when using the ‘lexstat’ scorer.
- Idtype: ‘loose’ or ‘strict’ manner to unify matches over words. Default is ‘loose’.
- Runs: number of runs for ‘lexstat’ scorer. Default is 1000. More runs increases precision.
- Store: directory to store analysis wordlist.
- Series: qualifier for file.

- Label: secondary qualifier for file.

Performs a Cluster or Partial analysis over the entire KeyPano dataset.

Stores the resulting analysis for reporting.

Report based on stored analysis

Using the analysis file as a base performs a few different reports over families, languages, concepts and words.
For some reports, uses the previous IDS indication of ‘borrowed’ word and the calculation of cognate crossing family boundaries to measure borrowed word detection

Reports:

- Summary of cognate id counts by family for cognates which cross family boundaries.
- Rough borrowed word detection using interfamily cognates as indicator of borrowed word and IDS borrowed word indicator.
- Detail report of words by detection status (fn, fp, tn, tp).

Arguments for report from analysis file:

- Store: directory from which to get the intermediate analysis file.
- Infile: name of intermediate analysis file.
- Family: Family of languages to report. Default is None. None reports on all families.
- Exclude: Family to exclude from putative recall, precision, F1 scores. Default is None. Typically Indo-European since they are frequent source of borrowing. This same family name is used to designate likely borrowing sources.
- Index: Index of threshold from analysis to use in report. Default is 0. When analyze uses multiple thresholds, this index selects which to use.
- Status: Which words to report for borrowed detection status. fn, tn, fp, tp, f, t, ntn, all.
- Output: Output directory for reports. Default is ‘output’.
- Series: Qualifier for report.

Calculate

Calculate the putative hierarchical relationship tree based on the intermediate analysis file.

Arguments for calculate:

- Dataset: File name for intermediate analysis file. Full file path including directory, name, file suffix.

Report Donor

Performs pairwise analysis versus selected donor languages [Spanish and Portuguese], and reports on potential cognate pairs, borrowed word and detection performance, individual borrowed word detection status.

Reports:

- Summary of pairwise alignment with donor languages.
- Summary of detection performance by language and by family based on alignment and IDS indication of borrowed word.
- Detail report of words by distance from candidate donor words.
- Detail report of words by detection status (fn, fp tn, tp, ...)

Arguments for pairwise alignment and report

- Model: 'sca' or 'asjp' sound class model. Default is 'sca'.
- Threshold: threshold limits below which cognate clusters are identified. Default is 0.4.
  - Higher is less restrictive and results in more false positives.
- Limit: Extended limit for reporting other results even if not below threshold. Default is None.
- Status: Which words to report for borrowed detection status. fn, tn, fp, tp, f, t, ntn, all.
- Mode: 'global', 'local', 'overlap' or 'dialign' methods of alignment. Default is 'overlap'.
- Donor: List of donor languages to consider. Default is ['SpanishLA', 'PortugueseBR']

- Output: Directory to write reports.
- Series: Report qualifier.