Topological Analysis of Syntactic Structures

Reference: https://arxiv.org/abs/1903.05181

During human history, languages appear, evolve, influence each others, and possibly be distinct. For example, English, Dutch, Icelandic and other languages in the Germanic language family are all evolved from the proto-Germanic. On the other hand, French, Swedish and those languages in the Romance language family are evolved from the ancient Latin. All the languages mentioned above belong to the Indo-European language family, all evolved from a single pre-historical language: the Proto-Indo-Europeans.

A language can be characterized via its lexical, morphological, phonological and syntactic parameters（词汇，形态，语音，句法）. Languages close in historical evolution are expected to be similar in these parameters. In this artical, we mainly focus on the last one: syntactic parameters of a language, and detect the historical evolution of world's different languages.

Data Structure

There are two data bases: SSWL (Syntactic Structures of World Languages) and LanGeLin , recording syntactic parameters of world languages, see page 8 and 10.

The SSWL data base consists of 116 binary parameters describing for example relations of adjectives to nouns and degree words, properties of numerals of a language and so on, ranging in $\{0, 1\}$. The LanGeLin database is similar, but some properties can be 'undefined', and hence are trinary parameters, ranging in $\{-1, 0, 1\}$ with 0 representing undefind data.

Each data point represents a language in high-dimensional Euclidean space, whose dimension in our cases are 116 and 91. The distance between data points is the Euclidean distance.

In our analysis we focus only on certain language families, such as the Indo-European language family, the Niger-Congo language family, the Austronesian language family and the Afro-Asiatic language family. We also consider the Ural-Altaic hypothetical family. See page 10-11 for the list of languages contained in the SSWL and LanGeLin databases.

Problem of the SSWL database: many languages are incomplete, due to the lack of information. Solution: only consider languages that are at least 50% complete, and only consider the non-absent parameters.

Main Problems

Three main problems:

- To what extend the persistent H_0 of the data can be used as an alternative method for the reconstruction of phylogenetic trees of language families?

- What is the meaning of persistent H_1 of the data from the point of view of historical linguistics?
- An estimation of dimensionality for different language families as a measure of how spread out the syntactic features are across languages in a given family.

Method to Use

We use the persistent homology for the data points to obtain the relatedness of the languages. The 0-th persistent barcodes give us information about persistent connected components, which further correspond to the phylogenetic tree of languages. The 1-st persistent barcodes correspond to cycles in the data, which imply some homoplasy phenomenon.

Data Pre-Processing

We use the PCA to reduce the dimension. The PCA (Principal Component Analysis) find the directions in high-dimensional Euclidean space that maximize the variance of the data, sorted in a decreasing order. Then the data can be regarded as mainly distributed on the subspace spanned by the first several directions (for example, the first 60%) maximizing the variance. Specifically, this is done by computing the covariance matrix of the data and then diagonalize it, obtaining the largest eigenvalues.

By finding appropriate directions, the PCA gives linear combinations of the original features. As a result, the binary or trinary features in the SSWL or LanGeLin databases become real-valued features after running PCA, lying in lower-dimensional Euclidean spaces.

The PCA is crucial since the model is unable to handle the persistent homology computation over the entire high-dimensional space.

The level of variance of the PCA will affect the final result.

Cluster Analysis

We construct the Vietoris-Rips complex from the data points after the PCA, then we compute the 0-th persistent barcodes. Using the following algorithm, we get a tree from the clustering.

The tree-constructing algorithm. Step 1, fill the empty spaces in the data with the midpoint of the data;

Step 2, Do PCA and take up to the percent variance we choose (for example, 60% or 80%)

Step 3, For each radius, construct the Vietoris-Rips complex from the data and get clusters $C = \sqcup_r C_r$.

Step 4, construct tree by regarding each C_r (for every r) as nodes and regarding C_i as a child of C_j if $C_i \subseteq C_j$ and there is no cluster C_k such that $C_i \subsetneq C_k \subsetneq C_j$.

The tree we get is much similar to the phylogenetic tree of languages, while there are still some differences. First, the data is not perfect so the closeness of languages is not as we expected. Second, the tree cannot imply historical evolution of languages. For example, older languages may be parents of younger languages using this algorithm.

The Indo-European language family. The Indo-European language family consists of most of European languages together with those on the Northern India subcontinent and Iranian Plateau. Popular languages in the Indo-European language family include English, Germanic and Dutch, which belong to the subfamily Germanic languages; France, Italian, Spanish and Portuguese, which belong to the subfamily Romance languages; Greek and its variant accents, which belong to the branch of Greek language family and so on. According the historical linguistic research, all these languages are original from the proto-Indo-Europeans.

See figure 23 for the persistent components tree for the Indo-European language family in the SSWL database. One can see that the syntactic similarities between the ancient languages is more closely clustered than that of their modern descendants. Most of languages are clustered correctly while several languages such as English, Icelandic and Faroese are misplaced. See also figure 24, the maximal number of non-trivial clusters in this case is about 10. These mistakes are not caused by the data in SSWL, since other methods can reconstruct the correct phylogenetic tree; instead, it may be caused by intrinsic weakness of our method or by the variance selection in the PCA.

See figure 25, 26 for the persistent components tree for the Indo-European language family in the LanGeLin database. With this database, English is correctly clustered to the family of west Germanic family while Icelandic is still misplaced, far away from the north Germanic family. Bulgarian is clustered as a singleton instead of with other Salvic languages, which is the correct place for it. The LanGeLin database groups all of the Germanic languages together within the same subtree.

Germanic Language Family. Germanic language contains the world's most famous language, English. All Germanic languages are derived from Proto-Germanic, spoken in Iron age of Scandinavia (including Sweden and Finland).

In the clustering tree of SSWL database, the Northern Germanic languages are closer to the Romance languages other than the west Germanic languages. Also English and Icelandic are misplaced. See figure 27 for the clustering tree for Germanic Language family of SSWL data. Except for the fact that the English sub-cluster joins the North-Germanic sub-cluster before nerging with the West-Germanic sub-cluster, this cluster does show the expected West-Germanic/North-Germanic split. Also see figure 28.

See figure 29, 30 for the clustering tree for Germanic language family of LanGeLin data. Except for the incorrect grouping of Icelandic with Old English, the clustering gives the West/North Germanic sub-clusters correctly. See page 56 for the comparison with different PCA variance and the correct phylogenetic tree.

Romance Language Family. The Romance languages are modern languages that evolved from Latin, consisting of Calabrian (an invariant of Italian), Catalan (co-official language of Spanish), French, Italian, various Italian dialects, Latin, Late Latin, Portuguese, Romanian, Spanish and so on.

In the full clustering of the SSWL database, the sub-cluster contains many of the Romance languages, while it does not contain French and Romanian. French and Old French are close to each other, but are clustered as singletons and are very far away from other Romance languages. It is known that Romanian shares grammatical features with non-Romance languages such as Greek, Bulgarian and Serbo-Croatian, which is indeed the case in the full clustering tree.

See figure 33 for the clustering structure of Romance Language family for the LanGeLin database. There is a singleton 12 corresponds to Latin, which should be the root of the tree. The two main clusters 20 and 21 splits the Romance languages precisely: cluster 21 consists of all southern Italic dialects while cluster 20 consists of all other main Romance languages.

The Hypothetical Ural-Altaic family. The Uralic languages include the Uralic languages Estonian (spoken in the Northeast of Europe), Finnish, Hungarian and local Russian languages Udmurt, Yukaghir and Khanty. Altaic languages consist of Turkish, Buryat (spoken in Mongolia), Yakut, Even and Evenki, as well as Japanese and Korean that were proposed earlier, although these two languages were later discarded from the Altaic hypothesis.

The Ural-Altaic is a linguistic proposal of uniting two language families Uralic and Altaic together. Our peristent components tree for the LanGeLin data seems to be supportive to this hypothesis. The tree places Japanese and Korean in closest proximity to each other, but very faraway from other Ural-Altaic languages. The rest of the Uralic and Altaic languages are all placed very closely together, see page 65.

Persistent First Homology

There are obvious differences between the persistent first homology of randomly-generated datasets of binary vectors and our language data. In the random data sets, H_1 appears for all sizes of clusters, while in the syntactic data from both the SSWL and the LanGeLin databases, nontrivial H_1 generators starts to appear only for bigger clusters. The barcodes for H_1 also behave different. For random data, the H_1 generators generally persist shortly while for syntactic data, they persist longer and are typically more sparse. This suggest that in our case, the H_1 structure is more persistent and less coincidental.

See page 76 for an algorithm finding nontrivial generators of persistent generators.

Gothic-Salvic-Greek circle indicates influences between the Greek languages and South Slavic languages. The nontrivial H_1 generator may also captures the syntactic influence of New Testament Greek on Gothic. While it is known that the Gothic influence in the Proto-Slavic borrowing was primarily lexical, there is an indication of morpho-syntactic borrowing as well.

Xiabing Ruan presented on 17 May 2021 at SUSTech Applied and Computational Topology Seminar.