World Academy of Science, Engineering and Technology
International Journal of Cognitive and Language Sciences
Vol:13, No:05, 2019

Syntax and Words as Evolutionary Characters in Comparative Linguistics

Authors : Nancy Retzlaff, Sarah J. Berkemer, Trudie Strauss

Abstract : In the last couple of decades, the advent of digitalization of any kind of data was probably one of the major advances in all fields of study. This paves the way for also analysing these data even though they might come from disciplines where there was no initial computational necessity to do so. Especially in linguistics, one can find a rather manual tradition. Still when considering studies that involve the history of language families it is hard to overlook the striking similarities to bioinformatics (phylogenetic) approaches. Alignments of words are such a fairly well studied example of an application of bioinformatics methods to historical linguistics. In this paper we will not only consider alignments of strings, i.e., words in this case, but also alignments of syntax trees of selected Indo-European languages. Based on initial, crude alignments, a sophisticated scoring model is trained on both letters and syntactic features. The aim is to gain a better understanding on which features in two languages are related, i.e., most likely to have the same root. Initially, all words in two languages are pre-aligned with a basic scoring model that primarily selects consonants and adjusts them before fitting in the vowels. Mixture models are subsequently used to filter 'good' alignments depending on the alignment length and the number of inserted gaps. Using these selected word alignments it is possible to perform tree alignments of the given syntax trees and consequently find sentences that correspond rather well to each other across languages. The syntax alignments are then filtered for meaningful scores—'good' scores contain evolutionary information and are therefore used to train the sophisticated scoring model. Further iterations of alignments and training steps are performed until the scoring model saturates, i.e., barely changes anymore. A better evaluation of the trained scoring model and its function in containing evolutionary meaningful information will be given. An assessment of sentence alignment compared to possible phrase structure will also be provided. The method described here may have its flaws because of limited prior information. This, however, may offer a good starting point to study languages where only little prior knowledge is available and a detailed, unbiased study is needed.

Keywords : alignments, bioinformatics, comparative linguistics, historical linguistics, statistical methods
Conference Title : ICELLET 2019 : International Conference on Evolutionary Linguistics and Language Evolution Theories
Conference Location : Montreal, Canada
Conference Dates : May 23-24, 2019