Abstract

The framework introduced in the Master Thesis of A. Varga to measure distances (similarity) between formal languages and between grammars based on distances between words is expanded. It is based on approximating languages by their finite subsets and using monotone sequences of such finite approximations to define an infinite language in the limit. Distances between finite languages are defined and extended to distances between monotone sequences of finite languages leading to distances between infinite languages. The framework captures several distances studied in the literature.

Context-free grammars with energy, enabling finite approximations emphasizing "syntactically important" parts of words are further investigated. Grammars with energy are used to extend distances between monotone sequences of finite languages to distances between context-free grammars. The "generating capabilities" and generation speed of grammars with energy are also examined.

A basic toolkit for monotone sequences of finite languages and distances between languages resp. grammars is provided. As part of this toolkit a non-symmetric version of distances is defined, providing additional characterisation of distances in general. Additional properties of distances between grammars are derived by restricting the "energy use" of grammars with energy.

A new distance, the asymmetric edit distance, is constructed based on the framework providing alternative characterisation of language intersections, while also showcasing the framework presented. Using the asymmetric edit distance language intersections are generalised and their relation to language neighborhoods is presented. Common subgrammars of context-free grammars are also studied in conjunction with the asymmetric edit distance.

Some methods of estimating the distances are presented to be used in cases where the distance is not computable or difficult to compute.

Keywords: grammars with energy, similarity of languages, distance between languages, finite approximation of languages, language intersection, context-free languages