An Application of Neural Nets to Comparative Linguistics

G. Sampath
Division of Computer Science and Mathematics
Marist College
Poughkeepsie NY 12601

Patterns of cognates in languages within a language group and of synonyms across language groups may be studied using different classes of neural nets. A distance (metric) may be defined (with varying degrees of complexity) over a set of languages based on corresponding patterns at different levels of language structure and function (where a pattern may be defined based on sound, or, for that matter, any reasonable, correspondences). One simple distance measure may be defined as the average Hamming distance between two output pattern sets as measured from one or more nets storing patterns in a basic cognate or synonym set. For a collection of languages, this leads to a spatial map in which the languages occupy regions (perhaps intersecting) in some n-dimensional space. In the next step, this basic model is enlarged by adding language-specific words (both open and closed) to the base so that the evolution of the map over that space may be studied and compared with conventional approaches to the study of language evolution as a function of migration over time and with different classifications of languages as currently advanced in comparative linguistics. The advantage of such an approach is the possibility of reconciling diverse theories from the viewpoint of pattern recognition and classification.

Work is presently under way in developing such a map, with languages being compared at the word level. The first step involves coding each word in some base set in terms of the phonological representation of its morphemes. The comparison is done by activating a neural net with the base morpheme set for a language pair that is a cognate set for 'related' languages and a synonym set for 'unrelated' languages. The class of nets initially chosen (mainly for its simplicity) is the Hopfield net. Thus a language map may be generated by activating a pair of Hopfield nets H_i and H_j for a language pair L_i-L_j with their base morphophonological code sets P_i and P_j respectively. On completion of the activation process these two code sets end up being stored collectively in H_i and H_j respectively. Next H_i and H_j are successively presented with the morphophonological code for each respective cognate (or synonym) and the output pattern of each net measured. If w_{il} and w_{jl} are two cognates (or synonyms), the resulting net outputs are p_{il} and p_{jl}, and the measured Hamming distance between p_{il} and p_{jl} is h_{ijl}, then the distance d_{ij} between languages L_i and L_j is defined as

$$d_{ij} = \text{average } \{ h_{ijl} \} \text{ over } P_i \text{ and } P_j.$$

These distances are then used to characterize a two- dimensional euclidean space by plotting a set of 'points' on the Cartesian plane using distance geometry techniques. The 'points' represent the set of languages $\{L_i\}$ being compared and really define a set of regions R_i of the plane some of which may overlap. These regions may then be used to define language groups and distances between them.

Experimental results in the form of a language map for a variety of languages will be communicated in due course.

Keywords: Comparative linguistics; Neural nets; Language classification; Pattern recognition