جامعـة نيويورك أبـوظـبي

NYU | ABU DHABI

N-to-1 Analysis of Linguistic Feature Patterns in the SSWL Dataset: An Expansion of Greenberg's Linguistic Universals through DataMining

Juan Felipe Beltran Capstone Project
Advisor: Dennis Shasha
BS Computer Science 2014

WHEN V2 01_Declarative Verb-Second IS neg AND Order N3 05_Adjective Demonstrative Noun IS neg
THEN Neg 09_Standard Negation is Reduplication IS neg

Motivation: Greenberg demonstrated a set of linguistic universals in his paper "Some Universals of Grammar with Particular Reference to the Order of Meaningful Elements." The universals make statement of this format:

1. "In declarative sentences with nominal subject and object, the dominant order is almost always one in which the subject precedes the object."

. . .

43. "If a language has gender categories in the noun, it has gender categories in the pronoun."

Given the recently published dataset of language features - hosted in the Syntactic Structures of the World's Languages Database (SSWL), the creation of such statements for any subset of languages, or the dataset as a whole becomes an approachable problem.

In their simplest form, Greemberg language universals are of the format:

```
IF A.feature.value = X
AND  B.feature.value = Y
. . .
THEN C.feature.value = Z
```

The Dataset: SSWL currently contains a notation of 93 different linguistic features over 213 languages. Additions to the dataset are incidental based on publication or new linguistic analysis.

The features are further separated into 10 different feature groups, depending on what aspect of linguistics the feature addresses. Features range from simple word-order patterns (Subject - Verb - Object, Subject - Object - Verb)to query-answer polarity connotations (Polarity-reversing answer by affirmative and special particle).

All documented features have one of three values "yes" when the feature is present "no" when the feature is not present and "NA" when the feature is inapplicable to the linguistic structure of the language. (For instance, a language without subjects cannot be said to have or not have Subject - Verb - Object word order).

Each language is listed along with the percentage of their Property:Value pairs that have been set (some features are yet to be documented). Language completion ranges wildly, from completely documented, to not at all.

The Analysis: After a heavy bout of data-parsing the dataset was simplified into 213 93-dimensional feature vectors, with the values 1, 0, and -1 standing for present, not present, and not applicable, respectively.

Feature expression for the given dataset is done in the most straightforward way, where all N- combinations without repetition are tested against every feature in every language. Once the iterator finds a contradiction to the temporary rule we've created, it stops and move on to the next combination. If no contradiction is found we classify this as a linguistic universal for the dataset.

In order to avoid redundant rules, we make sure that every no supersets of previously found rules are tested again. For instance, if we have found a 1-to-1 rule "IF A IS 1 THEN C IS 1" we will not test any combination that includes A against C for any subsequent run of the algorithm.

We see O(L*C(N,F)*F) efficiency where L is the number of languages, F is the number of features, and C is the combination without repetition function into groups of N features.

Output Parsing: Given the scale of the data, it is predicatable to find far more feature rules than are necessarily relevant to the field. In order to filter the results to maintain relevance the following rules were established:

(1) No rule will be presented unless the antecedents are of a different group than the consequent.
(2) No rule will be presented unless the number of languages explicitly demonstrating the rule are more than 10% of the total member count.
(3) No rule will be presented if it shows a relationship between "non-applicable" features.

Results which violate the above three rules, though true, are not very conductive to a better understanding of language. After filtering, we yielded no 1-to1 rules of relevance, and 6254 two-to-one rules of relevance.

Bibliography:

Greenberg "Some Universals of Grammar with Particular Reference to the Order of Meaningful Elements," Universals of Language, London: MIT Press, pp. 110-113.

Syntactic Structures of the World's Languages, "http://sswl.railsplayground.net/"