Summary 2: Producing Power-Law Distributions and Damping Word Frequencies with Two-Stage Language Models

Susan Cherry

This paper describes a two step process for developing statistical models that produce power laws. This framework is useful when analyzing the frequency of words in languages. Specifically, the framework contains both a generator step and a adaptor step. The generator step can be any probabilistic model and is the stage through which lexical items are generated. The second step is the adaptor stage during which the frequencies of the words are adapted to match the statistical patterns of natural language. The adaptor "damps" the frequencies of the words and "adapts" the word frequencies produced by the generator to fit a power-law distribution.

The paper specifically considers two adaptors: The Chinese Restaurant Process and the Pitman-Yor Chinese Restaurant Process. Both of these Processes are used in nonparametric Bayesian statistics. Next, the paper discusses relationships to other models and shows that the two-stage framework has three special cases: The Dirichlet-multinomial Model (equivalent to the TwoStage(CRP, Multinomial) model), the Dirichlet Process (equivalent to the TwoStage($CRP(\alpha), P_\psi$) model), and the Pitman-Yor Process (equivalent to the TwoStage($PYCRP(a, b), P_\psi$) model).

Different adaptors have different effects on frequencies. Using the CRP or PYCRP as adaptors is approximately equivalent to estimating parameters from either log transformed or inverse-power transformed counts. Overall, the paper explains why adopting this framework improves performance-it compensates for the rich-get-richer process that produces the power law distributions that are seen in natural language.

The paper concludes with two experiments that show how the PYCRP adaptor can be used in the two-stage framework. The results of the experiment suggest that partially damping frequencies might be more effective than fully damping frequencies. They propose two main reasons that using PYRCP adaptor to damp frequencies give better morphological segmentations that simply directly from corpus frequencies:

1. The generator model assumes that stems are suffixes are independent given the morphological class, but this is not always the case.
2. The most frequent words in any language tend to be irregular and due to the power low distribution, these words dominate corpus statistics.

The paper ends by suggesting that the two-stage framework shows how ideas from non-parametric Bayesian statistics can be valuable in computational linguistics. Defining models with infinite complexity allows the models to grow as more data are observed.