Breaking the Zipfian Barrier of NLP

Monojit Choudhury
Microsoft Research India,
Bangalore
monojit.choudhury@gmail.com

Abstract

We know that the distribution of most of the linguistic entities (e.g. phones, words, grammar rules) follow a power law or the Zipf's law. This makes NLP hard. Interestingly, the distribution of speakers over the world, content over the web and linguistic resources available across languages also follow power law. However, the correlation between the distribution of number of speakers to that of web content and linguistic resources is rather poor, and the latter distributions are much more skewed than the former. In other words, there is a large volume of resources only for a very few languages and a large number of widely spoken languages, including all the Indian languages, have little or no linguistic resource at all. This is a serious challenge for NLP in these languages, primarily because state-of-the-art techniques and tools in NLP are all data-driven. I refer to this situation as the "Zipfian Barrier of NLP" and offer a mathematical analysis of the growth dynamics of the linguistic resources and NLP research worldwide, which, afterall, is very much a socio-economic process. Based on the analysis and otherwise, I propose certain technical (e.g. unsupervised learning, wiki based approaches to gather data) and community-wide (e.g. acceptance of language specific works and resource building projects in top NLP conferences/journals, Special Interest Groups) initiatives that could possibly break this Zipfian Barrier.

Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages, pages 5–6,
Hyderabad, India, January 2008.