Intelligence Resource Collection for Low-Density Languages

Jaime Carbonell

Whereas there are over 6,000 languages, less than 100 have ample resources that are easily collected. The next tranche of several hundred languages may have significant monolingual resources, but very little parallel text used to train corpus-based machine translation. Beyond the top 1,000 languages even monolingual text is rather scarce. This talk addresses methods to: 1) build language technologies methods, in particular MT methods, that require fewer scarce resources, and to 2) collect resources dynamically via active and proactive machine learning, so that the most crucial information is elicited when collection/annotation costs are high or information sources are otherwise limited. We focus on emerging new paradigms for MT based on richer linguistic priors, and consequent collection of human judgments primarily from bilinguals (e.g. translations, alignments, corrections, or heavier linguistic annotation) to supplement small-scale existing collections that may exist for the languages of interest.