Addressing the Language Resource Gap through Alternative Incentives, Workforces and Workflows

Invited keynote lecture

Christopher Cieri

LTC 2017
Poznań, Poland

Abstract

After several decades of development, the dream of Human Language Technologies reaching a state of maturity that spawns products that benefit an enthusiastic user population seems upon us. Unfortunately, this is only true for some languages, only some technologies applied to those languages and only some linguistic genres for those languages and technologies. The advances in HLT have historically been fueled both by new algorithms and by language data; the same will be true for future advances.

For most languages, genres and technologies, the dearth of Language Resources relative to demand impedes progress, as it had done for the past several decades even though governments and companies around the world invest immense effort in creating requisite data. The Language Resource deficiency even affects languages with worldwide economic and political influence but for most of the world's 7000 linguistic varieties, the absence is acute. Current approaches cannot hope to meet the resource demand for even a reasonable subset of the languages currently spoken because they seek to document phenomena of great variability principally using resources, such as national funding, that are highly constrained in terms of amount, duration and scope.

While the struggle continues to squeeze increasing volumes of language data from shrinking budgets. A few initiatives scattered across HLT but a bit more commonly outside our communities have begun to find success employing new and different incentive models to attract new workforces and implementing appropriate workflows to ingest what they produce. This paper begins by describing efforts among HT developers and beyond, to augment the incentives of monetary compensation in order to elicit greater contributions of linguistic data, metadata and annotation. It also sketches the adjustments to workforces, workflows and post-processing needed to collect and exploit data elicited under novel incentives. Finally, it introduces a new initiative to more fully implement these techniques among international data centers that serve the HLT community in order to benefit researchers, educators and technology developers who work with language.