LREC 2014 Workshop

CCURL 2014
Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era

Reykjavík (Iceland), 26th May 2014

Oral Presentation

Title

Multilingual Lexicography with a Focus on Less-Resourced Languages: Data Mining, Expert Input, Crowdsourcing, and Gamification

Authors

Martin Benjamin, Paula Radetzky

Abstract

This paper looks at the challenges that the Kamusi Project faces for acquiring open lexical data for less-resourced languages (LRLs), of a range, depth, and quality that can be useful within Human Language Technology (HLT). These challenges include accessing and reforming existing lexicons into interoperable data, recruiting language specialists and citizen linguists, and obtaining large volumes of quality input from the crowd. We introduce our crowdsourcing model, specifically (1) motivating participation using a "play to pay" system, games, social rewards, and material prizes; (2) steering the crowd to contribute structured and reliable data via targeted questions; and (3) evaluating participants' input through crowd validation and statistical analysis to ensure that only trust-worthy material is incorporated into Kamusi's master database. We discuss the mobile application Kamusi has developed for crowd participation that elicits high-quality structured data directly from each language's speakers through narrow questions that can be answered with a minimum of time and effort. Through the integration of existing lexicons, expert input, and innovative methods of acquiring knowledge from the crowd, an accurate and reliable multilingual dictionary with a focus on LRLs will grow and become available as a free public resource.