

An Intelligent Search Infrastructure for Language Resources on the Web: Project Overview

Tim Baldwin, Steven Bird & Baden Hughes
The University of Melbourne
{tim, sb, badenh}@csse.unimelb.edu.au

Presentation Overview



- Aims, Context, Motivation
- Project Outcomes and Deliverables
- Collaboration
- Research Support

Project Aims



- Build a language aware search engine, using existing technologies and software components developed by the Chief Investigators; develop a new infrastructure for language-centric research
- Fundamentally change the way that users can search for language/linguistic information on the web

Existing Work



- Leverage existing work in
 - Description and discovery infrastructure in Open Language Archives Community (OLAC) (2001-)
 - Locating web content of “linguistic interest” using LangGator crawler (2004-)
 - Robust and scalable methods for language identification (2005-)
 - Digital repositories for language data (2002-)

Research Support



- Australian Research Council (ARC)
 - Mission to advance Australia's research excellence to be globally competitive and deliver benefits to the community
 - Funds pure and applied research, industry linkages, fellowships
- Special Research Initiatives (SRI)
 - Focus on collaborative and integrative projects, short term intensive funding
- ARC SRI (E-Research)
 - Specifically targeted at large scale, technologically enabled interdisciplinary collaboration
- Our SRI Grant AU\$100K/12 months (CY 2006)
 - Targeted at engineering integration effort
 - “Pump-priming” for larger long term research and development effort

The Web Context



- The web

- Large (>30B pages ?)

- Search engines

- Incomplete coverage of the web's content

- “Smart” approaches to crawling, indexing and ranking

- “Dumb” approaches to content interpretation

- End user information need fulfillment

- Low precision information discovery

- Even lower precision information discovery in “low density” content areas

- High linguistic diversity both in terms of users and content

The Australian Context



- Australia is the most ethnically and linguistically diverse country in the world
 - 26% of Australians born overseas
 - 38% of Australians speak language other than English as first language
 - 62% of Australians speak a second language
- Institutionalised support for multilinguality
 - Major broadcasters deliver content in >130 languages
 - >500 'community' media organizations

Motivation



- Strong use cases for multilingual information discovery and access in Australian context
- Linguistic and cultural diversity in the local context motivates the use of globally scalable solutions for efficiency reasons
- Many interesting research questions in this area
- Existing work connects to this agenda

Outcomes



- Core deliverables from this project are services of various kinds
 - Language Crawler
 - Metadata Creation Tools
 - Language Archive
 - Search Engine

Language Crawler



- Software for “language-centric” web content retrieval
- Extremely high precision for language data
- Emphasis on languages of scientific, economic and cultural interest to Australia
- Discovered content stored in a centralised repository with long term sustainability
- Highly parallelised, bandwidth intensive
- Builds on existing work (LangGator) by CI Hughes
- Timeframe: January – March 2006

Metadata Creation



- Each resource automatically classified across multiple dimensions including language and resource type identification
- Open Language Archives Community (OLAC) metadata created automatically using various machine learning and classification techniques
- Content of resources indexed and integrated for exploration
- Timeframe: April – May 2006

Language Archive



- Identified high quality resources (dictionaries, grammars) for endangered languages are archived to ensure ongoing accessibility
- Infrastructure provided by Australian Partnership for Advanced Computing / Australian Partnership for Sustainable Repositories
- Classified resources are used as seed data for iterations of the LangGator crawler suite
- Timeframe: July – September 2006

Search Engine



- Users able to enter keywords such as language names, locations, linguistic constructions of interest and discover resources of interest
- Geospatial interface for search and results display
- Round trip exposure of content back to broad coverage search engines via gateways
- Builds on existing work by CI Hughes in the form of the OLAC Search Engine
- Timeframe: October – December 2006

Status March 2006



- LangGator Content Acquisition Module
 - Crawler 1.0 acquiring content, 1.6 million URIs of “linguistic interest” from languages ranked 2000+ by number of speakers identified and retrieved
 - Total data on disk 1.8Tb.
 - Crawler 2.0 (distributed collaborative crawling) in development
- Description Module
 - Adapted DCDot to support core OLAC descriptions
 - Adapting UCR’s iVia Core to handle language classification and linguistic data type classification
- Language Archive
 - Customization of E-Prints software for language data
- Search Engine
 - Enhancements to existing OLAC Search Engine (geospatial similarity, full API)

Linkages



- National

- University of Sydney
- Commonwealth Scientific and Industrial Research Organization (CSIRO), Defence Science and Technology Organization (DSTO)
- Australian Institute of Aboriginal and Torres Strait Islander Studies
- State Library of NSW, State Library of Victoria, National Library of Australia
- Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC)

- International

- SIL International, Linguistic Data Consortium
- University of Washington, Stanford University, California State University, University of Pennsylvania, University of Maryland, St Louis University
- University of Edinburgh