Mining language resources from institutional repositories

Gary Simons  
*SIL International and Graduate Institute of Applied Linguistics*

Steven Bird  
*University of Melbourne and University of Pennsylvania*

Christopher Hirt  
*SIL International and Payap University*

Joshua Hou  
*University of Washington*

Sven Pedersen  
*Graduate Institute of Applied Linguistics*

Digital Humanities 2011, Stanford Univ., 19-22 June 2011
OLAC is an international partnership of institutions and individuals who are creating a worldwide virtual library of language resources by:

- Developing consensus on best current practice for the digital archiving of language resources
- Developing a network of interoperating repositories and services for housing and accessing such resources

Founded in December 2000

- Now has 45 participating archives
- Combined catalog of over 105,000 language resources
The project context

► OLAC: Accessing the World’s Language Resources

  ▪ Collaborative NSF grants awarded to the Graduate Institute of Applied Linguistics (Dallas, TX) and the Linguistic Data Consortium (U. of Pennsylvania)

► Some project outcomes

  ▪ OLAC Metadata Usage Guidelines
  ▪ Infrastructure of metadata checks and metrics to promote use of best practices among participants
  ▪ Faceted search service that exploits best practice
This catalog, developed by the Open Language Archives Community (OLAC), provides access to a wealth of resources, including details of archives, tools, and software, all in one place.

Search for language resources

Sort by:
- Possible Sorts: all

Browse by:
- Archive
- Online
- Language
- Language family
- Geographic region
- Country
- Linguistic type
- Linguistic field

Browse the OLAC records by Geographic region or by Language:

- English (3521)
- Spanish (2925)
- Yuracare (1383)
- Aleut (1125)
- Central Yupik (1116)
- Ocaina (678)
- Ahhtena (618)

26 search facets: 14 controlled + 12 freeform
Problem statement

▶ Tens of thousands of language resources are on the web but can’t be found with conventional search:
  ▪ They may be in the deep web behind search interfaces
  ▪ Languages are not uniquely identified by names alone:
    ▪ Ambiguous names, alternate names, historical names, translations of names — OLAC solves this with ISO 639-3

▶ Major universities now preserve the work of their faculties in institutional digital repositories
  ▪ Can we build a system to automatically find language resources in the catalogs of these deep web sources and enrich the metadata with precise language identification?
Methodology

1. Train a binary classifier to determine whether a metadata record describes a language resource or not.

2. Train a named entity recognizer to identify language names in a metadata record.

3. Use OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) to harvest Dublin Core catalog records from institutional repositories.

4. For each catalog record, if the classifier says it might be a language resource and the named entity recognizer identifies a language, retain the record and enrich the metadata with the ISO 639-3 code for the subject language.
The language resource classifier

- We used MALLET—Machine Learning for Language Toolkit (from UMass Amherst)—to train a maximum entropy classifier.

- Training data:
  - Required a large collection of metadata records that covered the full range of human knowledge and that were already classified as to the nature of their content.
  - We used a collection of over 9 million MARC catalog records from the Library of Congress that was deposited into the Internet Archive by the Scriblio project.
  - We used bag-of-words features extracted from the title and subject headings of each MARC record.
  - To label each record as a language resource or not, we mapped the Library of Congress call number onto “Yes” or “No” based on an analysis of the LC classification system.
We implemented a Python function that:

- Scans the title, subject, and description metadata elements
- Finds longest matches of known language names
- Returns most likely language(s) based on length of match and strength of name

Sources of name data:

- Library of Congress subject headings for individual languages mapped to the corresponding ISO 639-3 codes
- Primary names, alternate names, dialect names from download data at ethnologue.com/codes (minus names that coincide with common words in stoplists of major European languages)
- Translation of major language names into the major languages used most frequently in the institutional repository metadata
The OAI harvester was seeded with 459 base URLs

- Found by querying the UIUC OAI-PMH Data Provider Registry for all providers with the word “university” in their description
- The harvest yielded 5,041,780 Dublin Core metadata records

The binary classifier was applied to each harvested record

- Returns a number between 0 and 1 representing the probability that the resource is a language resource
- Evaluating the results of random samples in successive probability ranges showed the classifier to be reasonably valid
- A random sample of 500 records with \(0.001 < p < 0.01\) yielded no language resources, so all records below \(p=0.01\) were discarded
- This left 71,238 records that might be a language resource
Results: Evaluating the binary classifier

Probability returned by binary language resource classifier

Number of language resources in random sample of 100 records

- Total
- Specific
Next step: Filtering based on language identification

Which of the 71,238 possible language resources should be entered into the OLAC catalog?

Basic strategy:
- Apply the language name recognizer to each record
- If it finds any, accept that record and enrich the record with the most strongly identified language(s).
- Except: filter out records that meet criteria which are found to correlate highly with incorrect results (discovered after preliminary evaluation of performance)

Result: 22,165 records were accepted
The final filtering criteria

1. Reject if it is assigned the special code [qqq] for formal languages and language disorders
2. Reject if it is assigned more than 3 languages
3. Reject if it is not assigned a subject language
4. Reject if it is from a repository specializing in an irrelevant subject
5. Reject if Format describes it as a photo or a physical artifact
6. Reject if it has a probability lower than 3.0%
7. Reject if it is in a Roman script language without a stoplist
8. Accept whatever remains
This record found at eprints.lib.hokudai.ac.jp is enriched with 2 language ids: 1 wrong and 1 right

<olac:olac>
  <dc:creator>Nagayama, Yukari</dc:creator>
  <dc:date>2008</dc:date>
  <dc:identifier>http://hdl.handle.net/2115/39564</dc:identifier>
  <dc:language>en</dc:language>
  <dc:publisher>Slavic Research Center, Hokkaido University</dc:publisher>
  <dc:subject xsi:type="olac:language" olac:code="rus"/>
  <dc:subject xsi:type="olac:language" olac:code="alr"/>
</olac:olac>
## Final evaluation of resource classification

<table>
<thead>
<tr>
<th></th>
<th>Accepted by filter</th>
<th>Rejected by filter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actually a language resource</td>
<td>175</td>
<td>24</td>
</tr>
<tr>
<td>Not a language resource</td>
<td>47</td>
<td>467</td>
</tr>
</tbody>
</table>

- **Accuracy** = 90%  
  (how often it was correct)
- **Recall** = 88%  
  (how many of the true resources it found)
- **Precision** = 79%  
  (how many of the accepted resources are right)

- Manual evaluation of 1% random sample of all records
Final evaluation of language identification

- Manual evaluation of the 260 language identifications made in the 222 accepted records in the 1% sample

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct identifications</td>
<td>186</td>
</tr>
<tr>
<td>Incorrect identifications</td>
<td>74</td>
</tr>
<tr>
<td>Missing identifications</td>
<td>22</td>
</tr>
</tbody>
</table>

- Recall = 89% *(how many of the actual languages it found)*
- Precision = 72% *(how many of the identifications are right)*
Known problems

Inspecting incorrect identifications reveals the following:

- 35% due to short words in non-English metadata
- 16% due to names used as adjective of ethnicity or place
- 14% due to names (esp. dialects) that are place names
- 12% due to short words missing from English stoplist

Inspecting missing identifications reveals the following:

- 43% due to the weighting heuristics giving the highest weight to the wrong language name
- 33% due to the name used not being in the training data for the language name recognizer (e.g. a non-English name)
Sample discoveries

In the 1% sample, resources from 53 distinct languages were correctly identified, e.g.,
- English (31)
- Chinese (16)
- French (15)
- Japanese (13)
- German (10)
- Spanish (7)
- Latin (6)
- Dutch (5)

And these more exotic languages:
- Ainu
- Basque
- Faroese
- Frisian
- Gothic
- Inuktitut
- Marathi
- Navajo
- Tibetan
- Yapese
- Alutiq (Yupik)
- Alutor (Russia)
- Hawaiian Creole English
- Itonama (Bolivia)
- Middle High German
- Occitan
- Pitcairn English
- Tausug (Philippines)
- Toba Batak
Conclusion

► This approach has mined 22,165 presumed language resources from over 5 million resources held in 459 institutional repositories.

► The currently achieved rates of recall and precision are beginning to yield usable results.

<table>
<thead>
<tr>
<th></th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resource identification</td>
<td>88%</td>
<td>79%</td>
</tr>
<tr>
<td>Subject language identification</td>
<td>89%</td>
<td>72%</td>
</tr>
</tbody>
</table>

► However, a number of things can still be done to improve the results further.