A paper proposal submitted to:
Digital Humanities 2011, Stanford University, 19-22 June 2011

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

Language resources are the bread and butter of language documentation and linguistic investigation. They include the primary objects of study such as texts and recordings, the outputs of research such as dictionaries and grammars, and the enabling technologies such as software tools and interchange standards. Increasingly, these resources are maintained in digital form and distributed via the web. However, searching on the web for language resources is a hit-and-miss affair. One problem is that many online resources are hidden behind interfaces to databases with the result that only a fraction of these resources are being indexed by search engines (He and others 2007). Even when resources are exposed to online search engines, they may not be discoverable since they are described in ad hoc ways that prevent searches from retrieving the desired results with high recall or precision.

This paper describes work being done in the context of the Open Language Archives Community (OLAC) to develop a service that uses text mining methods (Weiss and others 2005) to find language resources located within the hidden web of institutional repositories. It then uses the OLAC infrastructure to expose them on the open web and make them discoverable through precise search.

1. The OLAC infrastructure

As set out in its mission statement, the Open Language Archives Community[1] is "an international partnership of institutions and individuals who are creating a worldwide virtual library of language resources by: (i) developing consensus on best current practice for the digital archiving of language resources, and (ii) developing a network of interoperating repositories and services for housing and accessing such resources."

With respect to best practices, the community has thus far focused on developing recommendations for the metadata description of language resources[2] so that they can be discovered in search with high precision and recall. The OLAC metadata format[3] is an extension of Dublin Core[4]—the dominant metadata standard in the digital library and World Wide Web communities (Bird and Simons 2004). To support the need for precise search, the community has adopted five specialized vocabularies[5] for use in describing resources: *subject language,* for identifying precisely which language(s) a resource is "about" by using a code from ISO 639;[6] *linguistic type,* for classifying the structure of a resource as primary text, lexicon, or language description; *linguistic field,* for specifying relevant subfields of linguistics; *discourse type,* for indicating the linguistic genre of the material; and *role,* for documenting the parts played by specific individuals and institutions in creating a resource.

With respect to the network of interoperating repositories, there are now more than 40 institutions that are sharing their language resource metadata to create a virtual digital library with over 90,000 holdings. Participating archives publish their catalogs in the XML format of an OLAC repository[7] and these repositories are "harvested" thrice daily by the OLAC aggregator using the Open Archives Initiative (OAI) Protocol for Metadata Harvesting[8] (Simons and Bird 2003)—another standard of the digital library community.

2. Mining for hidden language resources

The Open Access movement[9] has led to the widespread uptake of self-archiving of research results by university faculty and staff. It stands to reason that among the millions of resources deposited into open-access institutional repositories, there are thousands of language resources. But these resources are not typically accessible via general web search. This is because they are

[1] http://www.language-archives.org/
[2] http://www.language-archives.org/REC/bpr.html
[3] http://www.language-archives.org/OLAC/metadata.html
[4] http://dublincore.org/documents/dcmi-terms/
[5] http://www.language-archives.org/REC/olac-extensions.html
[6] http://www.sil.org/iso639-3/
[7] http://www.language-archives.org/OLAC/repositories.html
[8] http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm
[9] http://en.wikipedia.org/wiki/Open_Access_movement

hidden behind the search interfaces of hundreds of repositories and they lack precise identification as language resources. The question is, "Can we find the language resources in institutional repositories and then make them easy for the language resources community to discover?"

Our research addresses the problem by using text mining techniques. We have begun by training a binary classifier that identifies the likely language resources within an institutional repository. We used MALLET, the Machine Learning for Language Toolkit,[10] to train a maximum entropy classifier. For data to train the classifier we needed a large collection of metadata records covering the full range of human knowledge that were already classified as to the nature of their content. For this purpose we used the collection of more than 9 million MARC catalog records from the Library of Congress collection that was deposited into the Internet Archive[11] by the Scriblio project.[12] We used bag-of-words features extracted from the title and subject headings of each MARC record. To label each record as to whether it was a language resource or not, we mapped the Library of Congress call number onto the appropriate binary label based on a prior analysis of the Library of Congress classification system. The resulting set of 9 million training records was then given to MALLET to train a binary classifier for language resource identification.

The resulting classifier was applied to 5,041,780 Dublin Core metadata records that were collected by doing a complete harvest of 459 institutional repositories using the OAI Protocol for Metadata Harvesting. The list of base URLs to harvest was found by going to the University of Illinois OAI-PMH Data Provider Registry[13] and querying for all repositories with the word "university" in their Identify response. When applied to a metadata record, the classifier returns a number between 0 and 1 representing the probability that the resource is a language resource. This probability was added as a new metadata element to each harvested record. We then implemented an extension to the OAI-PMH interface on our metadata aggregator that allows us to request a ListRecords response of a given size that is a random sample of the records falling within a given probability range.

Figure 1 shows the result of evaluating the performance of the classifier by means of manually inspecting ten random samples of 100 records each representing the full range of probabilities assigned by the classifier. In the manual evaluation of the classifier results, each record was assigned to one of three categories: not a language resource, a resource about a specific language, or a resource about human language but no language in particular. Figure 1 plots the number of specific language resources found in each sample of 100 as the lower line;

[10] http://mallet.cs.umass.edu/
[11] http://www.archive.org/details/marc_records_scriblio_net
[12] http://about.scriblio.net/
[13] http://gita.grainger.uiuc.edu/registry/

the upper line adds the non-specific language resources. (Not plotted are a sample of 500 records for .001 < p < .01 in which were found 0 specific language resources and 4 non-specific resources, and a sample of 200 records for p < .001 in which 0 language resources of either type were found.) The graph demonstrates that the probabilities assigned by the classifier accord well with the actual proportions discovered by manual inspection, thus providing evidence for the validity of the classifier. The notable deviation from the expected trend is in the highest probability range. Inspection of the records in question showed that the majority of false positives were items from computer science about programming languages and formal language theory, leading us to hypothesize that the training data from the Library of Congress catalog was underrepresented in this area.

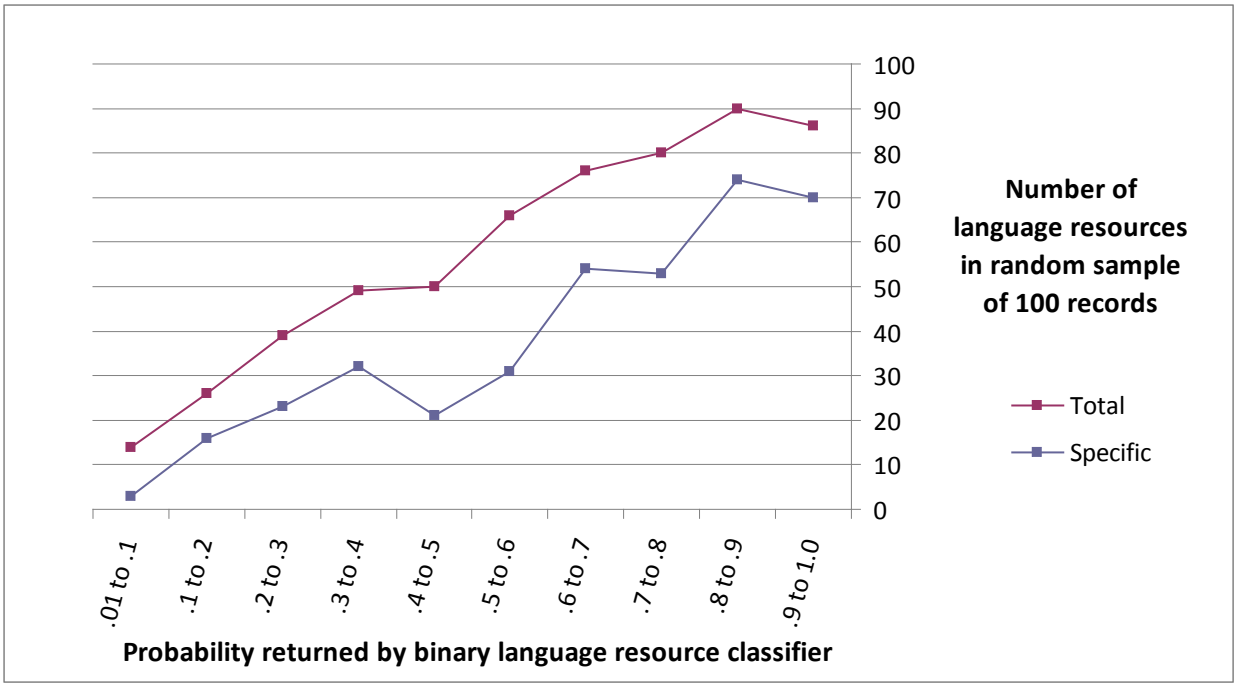

Figure 1: Evaluation of language resource classifier

Of the 4 million harvested records, only 52,000 indicate a probability greater than .01 of being a language resource. Multiplying the proportion of actual language resources found within each probability range by the total number of records falling within each range leads to the estimate that there are approximately 8,000 specific language resources within the set of 4 million harvested records.

3. Exposing the once-hidden resources

The next step in our research is to apply a multiclass classifier for language resource types to the metadata records for the 52,000 candidate language resources, as well as a named entity

recognizer for language names. The metadata records to which language resource type and language identification can be assigned with high probability will be enriched using the OLAC metadata vocabularies. They will then be entered into the combined OLAC catalog by creating a new OLAC data provider for these language resources that have been mined from institutional repositories. The final paper will report on the results of these efforts at metadata enrichment and show how the results are exposed to users through the two main OLAC services that support language resource discovery: an indexing service that provides a web page of relevant resources for each of 7,670 distinct human languages (as identified in the ISO 639-3 standard) and a faceted search service that makes it easy to find resources of interest by clicking on selected values of standardized descriptors to successively refine the search.

References

Bird, Steven and Gary Simons. 2004. Building an Open Language Archives Community on the DC Foundation. In D. I. Hillmann and E. L. Westbrooks, eds., *Metadata in Practice*, pp. 203–222. Chicago: American Library Association. <http://www.ldc.upenn.edu/sb/home/papers/mip.pdf>

He, Bin, Mitesh Patel, Zhen Zhang, and Kevin Chen-Chuan Chang. 2007. Accessing the deep web. *Communications of the ACM* 50(5): 95–101.

Simons, Gary and Steven Bird. 2003. Building an Open Language Archives Community on the OAI Foundation. *Library Hi Tech*, 21(2), 210–218. <http://arxiv.org/abs/cs.CL/0302021>

Weiss, Sholom M., Nitin Indurkhya, Tong Zhang, and Fred J. Damerau. 2005. *Text Mining: Predictive Methods for Analyzing Unstructured Information*. New York: Springer.