Dominic Widdows and Peter Lucas, MAYA Design, Inc.
{widdows,lucas}@maya.com

Pervasive Technology for Corpus-Based Research

Distributed Data, Search, and Collaborative Annotation

Though electronic media have already revolutionized the way large texts are stored and accessed, research still follows a 'one corpus, one location' pattern. Corpora are widely available (freely or under license) for search and download over the internet (examples include the British National Corpus [1] and the Project Gutenberg texts [2]), but such resources represent a very centralized approach to corpora. Even when corpora are in the public domain, current technology limits the data to single locations where the data can be accessed and searched, and annotation has to be done on a separate copy. Though there are distributed databases [4], this invariably means that *computation* is distributed, not collaboration between humans.

This paper presents an alternative distributed approach to using corpora. The technology is based upon the Information Commons architecture [3], a peer-to-peer 'universal database' where every object is represented by a scalable bundle of attributes and values indexed by a universally unique identifier. This information object is called a *u-form*.

Corpora are represented in u-forms using a `text_content` attribute. Other attributes can be added at will to provide metadata, such as syntactic and semantic annotation, relations to translated versions of the data, and (very importantly) attribution to sources and rights statements (data is not shepherded to other peers in the system without such attribution).

Indexing agents built against this peer-to-peer architecture enable search over widely distributed datasets. If the user wants to work more closely on matching items, their identifying keys can be collected to form a 'virtual corpus.' To improve performance the data itself can be replicated in the user's own repository.

Using the same peer-to-peer architecture, the user can annotate the corpus u-forms by adding extra attributes, and these annotations are shepherded back to the Information Commons. If the user does not have write-privileges to the original corpus data, they create metadata u-forms that refer unambiguously to the corpus material. Another user subscribed to the same interest-group automatically receives these annotations along with the special-purpose virtual corpus. Annotation efforts, that are all too often lost at the end of the projects that produced them, become available to future researchers in perpetuity.

The paper presentation will include demonstrations including distributed search, selecting material using drag-and-drop, and collaborative annotation using a 'back-of-an-envelope' interface. The system is a fundamental step forward in collaborative distribution, indexing, search, selection and annotation of freely available linguistic data.

References

[1] British National Corpus. `http://www.natcorp.ox.ac.uk/`.

[2] Project Gutenberg. `http://www.gutenberg.org/`.

[3] Peter Lucas and Jeff Senn. Toward the Universal Database: U-forms and the VIA Repository. Technical Report MTR02001, Maya Design, 2002.

[4] M. Tamer Özsu and Patrick Valduriez. *Principles of Distributed Database Systems*. Prentice Hall, 2nd edition, 1999.