Collaborative Annotation that Lasts Forever: Using Peer-to-Peer Technology for Disseminating Corpora and Language Resources

Fifth International Conference on Language Resources and Evaluation (LREC 2006).
Genoa, Italy, May 24-26, 2006.

Magesh Balasubramanya, Michael Higgins, Peter Lucas, Jeff Senn and Dominic Widdows
The authors are listed alphabetically, each author having contributed significantly to the technology described.

MAYA Design, Inc., Pittsburgh, Pennsylvania
{magesh,higgins,lucas,senn,widdows}@maya.com

Abstract

This paper describes a peer-to-peer architecture for representing and disseminating linguistic corpora, linguistic annotation, and resources such as lexical databases and gazetteers. The architecture is based upon a 'Universal Database' technology in which all information is represented in globally identified, extensible bundles of attribute-value pairs. These objects are replicated at will between peers in the network, and the business rules that implement replication involve checking digital signatures and proper attribution of data, to avoid information being tampered with or abuse of copyright. Universal identifiers enable comprehensive standoff annotation and commentary. A carefully constructed publication mechanism is described that enables different users to subscribe to material provided by trusted publishers on recognized topics or *themes*. Access to content and related annotation is provided by distributed indexes, represented using the same underlying data objects as the rest of the database.

1. Introduction

This paper describes the use of peer-to-peer technology for storing and disseminating linguistic information including corpora, syntactic annotation, language resources, and a publicly available gazetteer for universal georeferencing.

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 (www.natcorp.ox.ac.uk) and the Project Gutenberg texts (www.gutenberg.org)), but such resources represent a very centralized approach to corpora, and pale in comparison with the Web.

At the same time, voluntary or non-profit organizations such as the Internet Archive, Wikipedia, and Project Gutenberg have arisen, whose main contribution is the quality of the data they provide, not the physical storage. However, the current architecture of the Internet penalizes such organizations, because the greater their success, the greater their hardware and monthly bandwidth costs. On the other hand, current infrastructure encourages large private companies such as Google, Yahoo and Microsoft to become information owners as well as information service providers, since they can afford mass storage and bandwidth provided that these operations are part of a profitable business. This has already led to something of a crisis of confidence among some computational and corpus linguists. Should results obtained by "Google counts" be preferred over results obtained over much smaller traditional corpora? On the one hand, results that use a commercial internet search engine to provide corpus examples that match particular patterns inevitably give higher recall, and this is important in a field that values evaluation measures. On the other hand, such results are not replicable, and change not only with the growth of the internet, but with the daily business decisions of major search engines. In a scientific era that is seeing a resurgence in empirical methods, in fields that are deeply relevant to the future success of our civilization and the planet as a whole, such a basis for scientific decision-making fails traditional tests of acceptable rigor.

This paper proposes that peer-to-peer technology is a solution (and in the long term, the only viable solu-

tion) for these problems. But much more, this paper describes an existing implementation of such a solution, that provides the means for storing and disseminating corpus data, search tools that can be used to create special-purpose 'virtual corpora,' and semantic resources that can be used to provide persistently available annotation. We call this system the Information Commons. Some of the linguistic potential of the architecture was described in (Widdows and Lucas, 2005), and this paper describes the way in which many tools have been made available through this platform. We will discuss the storage and dissemination of corpora (and other datasets) in general, and also treat collaborative annotation in particular as an important problem that shows how the general features of our approach can be used to solve a difficult specific problem.

2. Requirements and a Solution for the Storage and Dissemination of Corpora

The general problem we face is to represent, store, and index corpora and other datasets in a manner so that the information is conveniently retrievable by a large number of users.

Information must be:

- Easy to find. Users must be able to query and create powerful indexes for discovering the information they need.
- Easy to replicate. A user must not have to depend on the availability of a centralized service. Similarly, we must not expect a single service to bear the expense of every user's needs.
- Easy to verify. Users must be able to associate information with a particular publisher, verify that the association is valid, and develop trust in a publisher's identity.

The first fundamental insight of the Information Commons approach is to represent every piece of information (whether it be a text, some multimedia document, or even an index) in a uniform way in a uniform identity space. Syntactically, each piece of information is represented as a *u-form*. A u-form is simply an extensible, typed bundle of attribute-value pairs with a universally unique identifier (UUID) (Lucas and Senn, 2002).

2.1. Indexing

To address the first requirement, ease of finding information, we depend on indexes.

It is important to note that we treat indexes as pieces of information like any other, and they are represented as u-forms. All we require from an underlying storage system is the ability to find a u-form given a UUID. More powerful search and retrieval strategies are handled by interrogating explicitly maintained indexes, not by relying on features of the storage layer. Moreover, we are free to create any style of index we need (we primarily use index structures modeled after B+-trees (Knuth, 1973) and R-trees (Guttman, 1984)). A detailed discussion of indexing in u-form space can be found in Balasubramanya et al. (2005).

2.2. Replication

UUIDs allow us to address the second concern: ease of replication. Each u-form is "named" by its UUID, not by a location-specific identifier (like a URL) or a database specific key. That means we are free to move u-forms around at will, and replicate them widely. If u-forms are modified by multiple agents, it is possible to introduce conflicting updates. A discussion of this complex issue is beyond the scope of this paper, except to say that the system we propose can detect and resolve conflicts, and the architecture we describe below for collaborative annotation effectively avoids conflicts.

2.3. Attribution and Verification

Finally, we must be able to verify that information is correctly associated with a particular author and publisher. Again, we desire to do this in a way that is independent of particular features or trust characteristics of the underlying storage technology. We do this by including attribution metadata with every u-form, along with digital signatures.

Every u-form may contain attributes describing its attribution. Our attribution scheme follows the Dublin Core standard (Dublin Core Metadata Initiative, 2004). A u-form may designate a publisher, one or more creators (authors), one or more sources, creation and modification dates, one or more languages, and one or more intellectual property rights statements.

It is not enough, however, for a u-form to claim attribution. This attribution must be verifiable. We support this by means of digital signatures (Schneier, 1996).

A writer in the Information Commons must possess a public-private key pair. If a writer creates or modifies any u-form, he must use his private key to sign the u-form. A reader can then use the associated public key to verify the signature. If the verification is not successful, then the reader can discard the forged u-form.

3. Corpora in the Commons

We have seen that u-forms provide a flexible information representation scheme that supports powerful information retrieval and information verification while supporting replication and making only very narrow demands on an underlying storage mechanism.

Individual texts are represented a u-forms with textual string values in a `text_content` attribute, and corpora are represented as collections of such u-forms. Additional attributes can be added at will to provide metadata, such as syntactic and semantic annotation, relations to translated versions of the data, and, as we have seen, attribution to sources and rights statements. For example, Figure 1 illustrates the way the first chapter of the novel *Moby Dick* is represented in a u-form.

Our information representation scheme and peer-to-peer technology combine to allow corpora to be replicated easily, and to allow third parties to maintain indexes and collections of subsets of documents. If, for instance, a user wants to work more closely on items matching some property, the UUIDs of the desired u-forms can be collected to form a 'virtual corpus.'

For example, the collection shown in Figure 2 contains UUID references to all verses from the King James Bible (i.e. the translation of Hebrew and Greek scriptures into English, made by the Anglican Church in 1611) containing the term *word*. This collection was generated by a simple keyword search query using a generic Information Commons search tool and a keyword index. For other theologians to study this collection (as they may wish to), they only need to replicate this collection of u-forms to a local storage system. For larger corpora, this becomes a significant cost saving, both in space for the researcher and in bandwidth for the network.

While we emphasize that the u-form approach is largely independent of the storage and retrieval mechanism chosen, we will later describe a particular peer-to-peer system that we have developed that makes it easy to replicate u-forms on demand and cache them locally.

Figure 1: Section of *Moby Dick* in the Universal Database. Attributes have been added to represent the author, publisher, rights statement, etc.

First, however, we describe the application of the Information Commons approach to collaborative annotation, a vitally important technique not only for linguistics researchers but for any users engaged in a rich discourse.

4. Requirements and a Solution for Large-scale Collaborative Annotation

Annotation of linguistic corpora can take many forms, such as part-of-speech tags, semantic annotation, and traditional footnotes and commentary. It is widely recognized that annotating a corpus is one of the main ways to add value to corpus data during the life-cycle of a project. However, with traditional inline annotation (such as XML markup), linguistic annotation can introduce at least two drawbacks. Firstly, the size of the corpus can increase dramatically, introducing extra cost to users who are not interested in this particular annotation. Secondly, a researcher may not have write-access to the original corpus data (systems that allow anyone to modify the text scale poorly and are susceptible to vandalism).

For these reasons, the need for "standoff annotation"

Figure 2: Virtual Corpus of verses from the King James Bible containing the term *word*

tools is gradually becoming recognized, and standoff annotation tools are becoming more widely available through platforms such as GATE (Cunningham et al., 2002). With standoff annotation, extra data is added to the corpus in separate data-structures, which contain pointers to the parts of the corpus they are intended to annotate.

If a standoff annotation mechanism is to be practical for large diffuse groups of commentators, it must have the following properties. It should allow:

- Reflexivity. There should be no fundamental difference between a text and an annotation. We must be able to annotate an annotation.
- Attribution. The author and publisher of an annotation should be clear. Trust depends on stable identities.
- Verifiability. Corpora and annotations must not be able to be forged.
- Storage independence. Corpora and annotations must not depend on the details of a particular storage or information retrieval system. This is necessary for future-proofing and ease of replication of data.

The Information Commons offers a mechanism for universal annotation that satisfies these requirements. We have already seen how the Commons provides attribution, verifiability, and storage independence for data in general and corpora in particular. Let us address annotations in particular and see how the Commons naturally handles the requirement of reflexivity.

4.1. Details of a Collaborative Annotation Mechanism

The Information Commons approach offers reflexive standoff annotation with verifiable attribution, *and* a convenient peer-to-peer storage, retrieval, and replication system. It is important to note that the storage infrastructure is independent of (though complementary to) the information representation scheme. This allows the two to evolve independently and assures that, as information technology becomes more advanced, corpora and annotations within the Information Commons stay stable and relevant. To clarify the presentation and emphasize this independence, we will discuss the details of the information representation scheme separately from those of the storage system.

4.1.1. Basic Information Representation

As we have seen, the fundamental unit of information in the Information Commons is a *u-form*. All we require from an underlying storage mechanism is the ability to look up a u-form by UUID.

Armed with u-forms, we make the following definitions:

- A *text* is a u-form representing the subject of annotations. Because u-forms are very expressive, a text may contain media other than the written word, though for linguistic purposes written texts are common. Any u-form may, in principle, be a text.
- A *corpus* is a u-form that links together many texts. Since a corpus is a u-form, it can be considered a text. Thus a corpus can be recursive: a collection of books that is each a collection of chapters, for instance.

- A *publisher* is an agent (person or organization) that vouches for an *author*. An author is the creator of a text. An author may be her own publisher, but a publisher may vouch for many authors.
- A *theme* is a u-form representing an area of interest. A theme may be as simple as a label, or may have a complex hierarchical structure that includes sub-themes. Anyone can create a theme at any time: communities of practice must agree on the meaning of themes. Because it is a u-form, a theme may be considered a text.
- An *annotation* is a u-form or set of u-forms that comments on a text with respect to a set of themes. The set of themes may be empty, or a singleton. Because it is a u-form, an annotation may be considered a text, and thus the subject of further annotations.
- An *index* is a u-form or u-forms that provides a mapping from a set of keys to a set of UUIDs (and hence u-forms).

To make standoff annotation practical, we must be able to define a relation over texts, publishers, themes, and annotations. In particular, a reader, given a text X, a publisher P, and a theme T, must be able to find some set of relevant annotations α.

Readers should also be able to find all themes available from a given publisher for a particular text.

A writer, similarly, must be able to produce a new annotation on a theme, and publish it with the aid of a publisher (perhaps himself).

To accomplish this, each publisher maintains an index u-form whose keys are the UUIDs of texts. The values in the index are UUIDs pointing to u-forms storing annotations sorted by theme. This allows a user to efficiently find the themes available for a particular text and publisher, and efficiently filter those annotations by theme.

A writer simply creates a new annotation on a particular theme, and then asks a publisher to associate his annotation with a text. If a writer is self-published, he maintains his own index.

4.1.2. Attribution and Verification of Texts and Annotations

We rely on the basic attribution and signing infrastructure described in section 2.3.

In the annotation scheme described in the previous section, authors create texts, annotations, and themes. Publishers create and maintain the indexes that associate annotations with texts and themes. Authors, thus, sign their texts, whereas publishers sign the indexes. By virtue of placing an author's annotation in an index, the publisher is vouching for the author. Publishers are free to pursue any policy: some publishers may be very conservative, only vouching for a small set of authors. Others may be much more cavalier. Some publishers may only vouch for an author in certain contexts. There is no single proper policy: it depends on the goals of the individuals involved.

It is important to note that the system permits anonymity: it is possible to create a u-form that is unsigned. We believe, though, that anonymous speech is often considered untrustworthy and likely to be discounted in many contexts. It is also possible to be pseudonymous: it is not necessary that a particular public-private key pair be easily traceable to a person's real-world identity. Again, though, pseudonyms may be considered by some to be less trustworthy *a priori*.

5. Peer-to-peer Storage and Retrieval

As we have seen, the u-form approach exploited by the Information Commons is largely independent of any particular storage system. It could be implemented using any number of traditional information storage and retrieval systems. All that is required is the ability to find a u-form given a UUID. All other content relationships and indexing are expressed in the u-forms themselves, so they are independent of the capabilities of the storage mechanism. However, many of our goals are better served by certain kinds of storage systems.

Traditional client-server systems are well understood and easy to build. But they suffer from a number of drawbacks. They are expensive to scale to many users (either readers or writers). They offer a single point of failure, so lack robustness. They are usually under the administrative control of one organization, which will often be unwilling to store and serve competing points of view. They are easy for powerful institutions to identify, censor, and control.

Truly large-scale robust information systems must be able to survive the brittleness of a client-server model. Peer-to-peer systems offer an attractive alternative. Current academic literature suggests that peer-to-peer systems can easily scale to millions of users.

Important systems include Pastry (Rowstron and Druschel, 2001), Chord (Stoica et al., 2001), Oceanstore (Kubiatowicz et al., 2000), and CAN (Ratnasamy et al., 2000). A number of systems enjoy wide current deployment, including Kademlia (Maymounkov and Mazieres, 2002) and BitTorrent (Cohen, 2003).

Because our information representation scheme does not depend on database specific keys, or location specific identifiers (such as URLs), it is easy to migrate from a client-server model to a peer-to-peer model. Most peer-to-peer systems offer only primitive search capabilities: typically only the ability to retrieve a single value for a key. Fortunately, this is all our design requires: search is supported by explicit u-form indexes rather than by features of the storage system.

Currently, the Information Commons relies on a peer-to-peer system we have developed called Shepherds (Lucas et al., 2005). It offers flexible topology management, optimistic replication, and lazy reconciliation. Our system currently stores millions of u-forms replicated across a number of sites.

6. Universally Available Resources for Semantic Annotation

One of the hallmark benefits of using the Information Commons architecture is that researchers have automatic access to all sorts of other datasets (and services) that are already part of the Information Commons. This occurs because UUIDs offer a single reference system that is domain independent, and the underlying storage mechanism makes replication transparent.

Examples currently include:

- Semantic resources such as WordNet (Fellbaum, 1998).
- The Information Commons Gazetteer. A publicly available dataset of over 5 million populated places (with names in over 20 languages) and geopolitical subdivisions down to the ISO first-level subdivisions. These have been gathered and fused from publicly available sources, so that they can be reproduced and used by researchers for free. This comprehensive resource is described by Lucas et al. (2006).
- Search tools (including automatic options for prefixed and stemmed searches). This means that researchers using the Information Commons as a platform can automatically search and cite one another's corpora. The annotation system leverages the standard search tools.

The power of the Information Commons becomes most clear when these resources are used in combination. For example, the Information Commons Gazetteer and the free search tools can be combined to give a free geo-referencing service. The combination of semantic resources such as WordNet and universal standoff annotation enables researchers to generate automatic semantic annotation for their corpora, without even downloading and installing WordNet!

7. Demonstrations

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

8. References

Magesh Balasubramanya, Michael Higgins, and Dominic Widdows. 2005. Distributed indexing and search methods for u-forms. Technical Report MAYAVIZ-05005, MAYA Design (prepared for MAYA VIZ).

Bram Cohen. 2003. Incentives build robustness in bittorrent. `http://citeseer.ist.psu.edu/cohen03incentives.html`.

H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. 2002. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*.

Dublin Core Metadata Initiative. 2004. Dublin core metadata element set, version 1.1: Reference description. `http://dublincore.org/documents/dces/`.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge MA.

Antonin Guttman. 1984. R-trees: a dynamic index structure for spatial searching. In *Proceedings of SIGMOD*, pages 45–47.

Donald Knuth. 1973. *The Art of Computer Programming*, volume III, Sorting and Searching. Addison-Wesley.

John Kubiatowicz, David Bindel, Yan Chen, Patrick Eaton, Dennis Geels, Ramakrishna Gummadi, Sean

Rhea, Hakim Weatherspoon, Westly Weimer, Christopher Wells, and Ben Zhao. 2000. Oceanstore: An architecture for global-scale persistent storage. In *Proceedings of ACM ASPLOS*. ACM, November.

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

Peter Lucas, Jeff Senn, and Dominic Widdows. 2005. Distributed knowledge representation using universal identity and replication. Technical Report MAYA-05007, MAYA Design.

P. Lucas, M. Balasubramanya, and D. Widdows. 2006. The information commons gazetteer: A public resource of populated places and worldwide administrative divisions. Genoa, Italy, May 24-26, May.

P. Maymounkov and D. Mazieres. 2002. Kademlia: A peer-to-peer information system based on the xor metric.

Sylvia Ratnasamy, Paul Francis, Mark Handley, Richard Karp, and Scott Shenker. 2000. A scalable content addressable network. Technical Report TR-00-010, Berkeley, CA.

Antony Rowstron and Peter Druschel. 2001. Pastry: Scalable, decentralized object location, and routing for large-scale peer-to-peer systems. *Lecture Notes in Computer Science*, 2218:329.

Bruce Schneier. 1996. *Applied Cryptography*. John Wiley and Sons, 2nd edition.

Ion Stoica, Robert Morris, David Karger, Frans Kaashoek, and Hari Balakrishnan. 2001. Chord: A scalable Peer-To-Peer lookup service for internet applications. In *Proceedings of the 2001 ACM SIGCOMM Conference*, pages 149–160.

Dominic Widdows and Peter Lucas. 2005. Pervasive technology for corpus-based research: Distributed data, search, and collaborative annotation. In *6th Conference of the American Association of Applied Corpus Linguistics*, Ann Arbor, Michigan.