# A Framework for Cross-Document Annotation

### David Day, Alan Goldschen, John Henderson

The MITRE Corporation 202 Burlington Road, Bedford, MA 01730, USA http://www.mitre.org/technology/nlp {day,alang,jhndrsn}@mitre.org

#### Abstract

We introduce a cross-document annotation toolset that serves as a corpus-wide knowledge base for linguistic annotations. This implemented system is designed to address the unique cognitive demands placed on human annotators who must relate information that is expressed across document boundaries.

### 1. Introduction

We are interested in building and evaluating systems that extract information about individuals and events from different textual sources. By operating over multiple documents, multiple facts about individuals can accrue to a single representation of that individual (or event), enabling a more complete description of the entity, and supporting evidential reasoning. In addition, maintaining records about the source of each piece of contributing information can provide a mechanism to cluster documents using domainspecific semantic relationships.

However, most of the linguistic phenomena that are routinely annotated in the corpus-based computational linguistics community have a relatively small locus of occurrence, almost always limited by the scope of the document itself. Within the sub-community that concentrates on information extraction, the layers of annotation range from the lexeme (segmentation, morphological analysis, part-ofspeech tagging), to simple short phrases ("named entity" tagging, "nominal entity" tagging, and similar tasks), sentences and their syntactic sub-structure, coreference relationships among phrases, and, at the upper extreme, "entity" and "event" tagging within the scope of a single document (herein document-level). In the latter case, the entities and events can be derived from linguistic cues occurring virtually anywhere within the document. It should not be surprising, then, that current tools are poorly suited to handle the challenges of cross-document annotation.

The same arguments for the utility of document-level information extraction apply equally strongly in the crossdocument case: we want to build systems that can reliably extract unique pieces of information from a text stream, independent of how that information is stated within the text stream. By operating over multiple documents, more facts about individuals and events can accrue to a single representation of that individual or event. In addition, there are distinct motivations for maintaining information about which specific documents information has appeared in:

- Such information provides a mechanism to thread documents, potentially at a finer granularity than simply by topic.
- Documents can be sorted and organized by the semantics of the extracted information, e.g., the temporal or-

dering of events, or other logical relationships among individuals or events (e.g., those that happened at the same location, etc.).

- Changes of authors' hypotheses and points of view can be better tracked across both time and source.
- Multiple sources can be compared and combined to increase or decrease support for extracted information, either relative to the language processing capabilities, or relative to models of the domain.

### 2. The Annotation Problem

Consider the following four selections of text, which we will assume derive from distinct documents (for example, from stories in different electronic news wire sources on different days).

- 1 ... US Ambassador Bill Richardson traveled to France today to begin the negotiations he requested with...
- 2 ... U.S. negotiator Richardson is not expected to discuss European Union trade concerns on his trip to meet President Chirac...
- **3** ...American farmers like B. J. Richardson are enthusiastic about the new hybrid seeds, in spite of European concerns about their potential for...
- 4 ... the president of France remains concerned about American agri-business and its willingness to...

Figure 1: Example text from four different documents.

In considering the problem of tracking identical people across their mentions in different news stories, we might represent the person entities mentioned in the documents as shown in Figure 2.

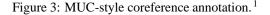
In theory, an intra-document coreference representation in which text spans are labeled with equivalence class identifiers can be used equally well for representing cross-document coreference relationships. This is the standard practice in the coreference annotation and evaluation community, with varying realizations. For example,

Person-1	Names	Bill Richardson	1
		Richardson	2
	Titles	US Ambassador	1
	Descriptors	U.S. negotiator	2
		he	1
Person-2	Names	Chirac	2
	Titles	President	2
	Descriptors	president of France	4
Person-3	Names	B. J. Richardson	3
	Descriptors	American farmer[s]	3

Figure 2: One way of depicting the information captured from four different example articles.

the Sixth Message Understanding Conferences evaluation (muc, 1995) of coreference used SGML annotations that used pair-wise links between annotated spans of text, as shown in Figure 3. This style of annotation can be depicted graphically as in Figure 4. The logical treatment of these pair-wise links is that a set of mutually coreferring expressions is represented by the transitive closure of all the constituent pair-wise links. Thus, while there are actually many different SGML element identifiers that may appear in any chain of coreferring expressions, one can view the annotations in one chain as identifying the individual mentions as a member of a particular equivalence class.

```
...<COREF ID="1">Bank Austria AG,
<COREF ID="2" REF="1">Austria's
largest bank,</COREF></COREF> said
<COREF ID="32">a covert,<COREF
ID="3"> 500-million schilling</COREF>
deposit</COREF>...
```



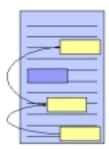


Figure 4: A graphical depiction of the coreference arcs that are annotated following the MUC6 annotation standards.

In practice, however, as one moves into the problem of identifying coreferring expressions across document boundaries, identifying coreferring pairs of text spans is untenable from the perspective of the human annotator. Consider the case in which an annotator working on annotating the 1 000-th document in a collection desires to coalesce a mention of Joe Smith (a very common name) with the previously annotated portion of the document collection. First, the annotator must find another annotated document containing a mention that could be this person. Then, if uncertainty remains, the annotator must determine if it is the same person by inspecting that document. Eventually, a mention of the correct underlying entity is found and the current mention is annotated with the corresponding equivalence class identifier.

The annotation problem is thus revealed: using this model, the annotator must repeatedly search the previously annotated documents, or keep discriminating information *in mind* to determine which mentions should be attributed to which underlying entity. This places an unreasonable burden on the annotator, and would severely hinder the ability to distribute the annotation task among different annotators.

### 3. Solution Framework

A framework that departs from traditional linguistic annotation is adopted to address this problem. The goal is to enable the annotator to accurately and rapidly find previously annotated candidate mentions and facilitate inspection of those mentions in context.

Standoff annotation in a database, together with a schema providing descriptions of searchable keys, provides a mechanism for finding candidate mentions. The annotator can query the database more precisely than using text scanning tools on large document collections. The annotator cannot be expected to predict which descriptions will be useful as keys for later search, however. Therefore, each mention in the database must have an associated pointer into its location in the document collection to enable viewing the mention in its context. An information retrieval open query mechanism using cosine distance in the vector space model should be utilized if specialized queries fail. Finally, text retrieval tools of the most primitive type, involving queries composed of string matching predicates can be used as a method of last resort.

#### 3.1. From Links to Free Standing Entities

Coreference as a linguistic phenomena is grounded not in the referring expressions on which coreference chains are constructed, but in the "entities" that are being referred to in the minds of the participating language users. So while it is relevant to represent and evaluate the ground expressions which give rise to coreference, a rich model of coreference can benefit from directly modeling the entities themselves. Thus, the next step for annotating and evaluating coreference is to explicitly capture the entities to which the coreferring expressions are referring (as depicted in the transition from Figure 4 to Figure 5).

The transition from "within-document" coreference to cross-document coreference introduces a model something like that in Figure 6. Here, the coreference links from entities in one document to entities in other documents are mediated through the cross-document repository. Information in the repository retains pointers to all the individual mentions within the distinct documents, but the default view from any single document is to a (single) entity in the

<sup>&</sup>lt;sup>1</sup>The annotations have been simplified by removing information indicating both the type of coreference relationship (IDENT, PART OF) and an indication of the head or MIN (minimum permitted) constituents of a particular phrase.

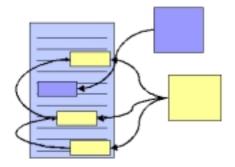


Figure 5: Entity-centric model of coreference annotation.

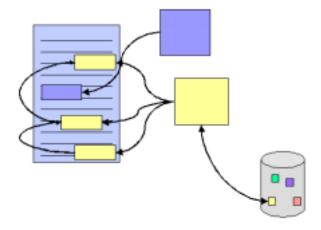


Figure 6: Entity repository for capturing entities and their references across a document corpus.

repository. The repository serves to present a unified, "coalesced" view of the entities which are defined and referred to in a multiplicity of ways.

### 4. Cross-Document Annotation Toolset

We have developed a set of tools that allow users to annotate a broad class of cross-document information using this approach. A single *entity repository*, implemented as a network service, maintains the corpus-level information being tracked. Individual documents within the corpus are annotated with pointers to the entity repository, and the repository, in turn, maintains references to all of the documents (and locations within documents) where information was individually annotated.

A range of annotation types can be supported through the specialized interfaces of the document annotation application, the Alembic Workbench (Day et al., 1997), whose results are stored either as SGML tags embedded within the source documents themselves or else as stand-off XML annotations using the emerging ATLAS annotation framework (Bird et al., 2000). For our current research we we are interested in annotating document collections with information about coreferring expressions that describe people, places, organizations, and simple relations among these entities (such as who works for whom, where an organization is located, etc.).

From the user's point of view, there are three

major components: (1) the document-level annotation application; (2) the cross-document entity repository query/browser application; and (3) a general purpose document browser. A depiction of how these elements might appear on a screen are shown in Figure 7. In this view the document-level annotation application actually consists of two elements of the Alembic Workbench-a text viewer and an entity viewer. In the lower left corner is a view of the source document text, with various text spans annotated and displayed with user-defined colors. In the lower right of the screen is the Alembic Workbench "relations" editor. This interface enables users to define attribute/value structures that can be associated with spans of text, as well as with other types of values (including user-entered strings, selections from pre-defined lists of values, and pointers to other entities of specified types, either in this or a different relation table).

In the upper left corner of the screen is the query interface to the cross-document repository. This interface is the front end of an application that runs on the client machine and sends user queries to the entity repository server somewhere on the network (using network protocols). The results of a query (those entities that match the query constraints) are returned and displayed in a different pop-up screen (not shown here). These returned values are in the form of whatever entities the server has been configured to store. (Entity structures are defined via an XML DTD, and data can be saved and restored via XML files.) The query and result interface components are in the process of being incorporated directly into the Alembic Workbench application to enhance their interoperability.

The goal of the users' interactions with the repository query facility is to explore the database in search of entities that are most likely coreferential with an documentlevel entity being annotated. When annotating a portion of text within the document, the user can signal that it should be added to the repository. All document-level annotations, whether unique or destined to be bound together with other entries in the repository, are added to the repository as atomic entries. Collapsing entries in the repository (for example, those that are deemed to be coreferential) into an equivalence class is a separate operation that does not destroy the initial atomic entry, but merely associates one or more entries in a way that is supported by the query and display mechanisms. In this way subsequent processing of the repository is free to associate and disassociate entries from one another without losing track of the separate pointers into the source documents from which this information was extracted.

The repository itself supports four major operations:

- Add Inserts new entities into the repository server, including possible substructure. For example, person entities might consist of names, titles, and descriptive phrases). Each entry in the repository always includes bookkeeping information, such as the original document and location within the document from which a piece of information was derived.
- **Query** Finds entries in the repository that satisfy various search criteria.

A CONTRACTOR			a second	Indiana Manageria			Statis error		
	File Edit	View Cla Ca	ommunicat	tor			He		
- queryclient. Ici	4.	2 3 4	き曲	468					
Label Durry String	👷 Bookna Aa 💰 Location: 🖬 🚛 / attai/est/geoglect/boopanil/bitst/fiela/BDD/histol 🖉 🐌 What's Pola								
(Hava) Shith	TTES / ME / DSD / WITHE Messanger / NLPQWITHE / Phone®s / MTP / Attaints / Yahoo /								
(Title) Pres. F Case Sensitve Match Search Clear Quit	thursday, 1 gang. accested fo attempt to	ut police qu the abductor or allegedly kill Pieside Treating in	ickly tr a deamad taking p ot Eduar two car	ncked down and sur led that the govern out in the list we d thermologies s. the bidesupers	ment free neven people				
Alexale: Verklesch ecticorpora2bdBidataEDTimini-corpus-2000-01-07Jcookedinewswine/BAPW19	neoral dist	and the host	ters 112	paline encircled o holed up in the inites) every. Gear	villege of gim'o Interior				
File Tag: Mention Options Utilities	ielp 📲	l			easty				
Alembic Workbench Charset: Latin-1 The MITRE Corpor	Relation: PILOTENTITY								
III III. Georgia (AP) Sideappers solved four United Nations military Electronic in the former South (publications) on Thurshop, but palice quickly tracked down and surrounded the anned		HameMoni STRINGE Bin		Hominathlentian STRINGFILL O n	PronounMention STRENGFILL 0 n	Type SETTE 11	- 1		
Thursday, but police guickly tracked down and surrounded the armed		6		Bussian forces		PER30	N		
DEED: The CONSTRUCT demanded that the <u>Servicement</u> free seven <u>Decem</u>		6 Abkhaut	n th	te province		GSP			
attempt to kill President Eduard Shevardnadre		14		prvenwood forces		PERSO	N		
Traveling in two cars, the sidnoopers abducted the dosenvers and stole one of their vehicles in the western descript zown of Ziggdid.		13		separatists The separatists		PERSO	•		
Several hours later, police encircled a house where the		8			one of the assailants	PERSO	N X		
Dimappers and the instances were holed up in the village of Dehishastari. 20 kilometers (12 miles) away. Georgia's Interior		New Rew		doction Delete Pr	About Relation	t Help	Dismiss		
Minister Kakha forcorroller said. Security efficient placed amored military vehicles outside the	Gell Cardents					Entarge	Shrink		
name and began negotiations with the analyzers, the interior									
The kidnep SETTS included two Storaway from Unuguey, 205 from Sweden and 505 from the storah (republic) along with their Georgian SETTS according to the SET along	Z								

Figure 7: An example layout of the component tools that can be used to annotate coreferring expressions across documents. (See text for further description.)

**Merge** Performs a reversible unification of two entries, enabling them to be viewed by repository clients as if they were a single entity (e.g., during queries, etc.).

Split Separates previously merged entities.

The contents of the server are accessible from a multithreaded XML database application whose internal knowledge representation (a Document Object Model) can be modified by re-defining the controlling DTD. We currently have defined a DTD that supports two kinds of entities: persons and organizations, both of which incorporate a small number of internal fields (including names, titles, and descriptors). Document-level annotations can be added directly to the entity repository via the repository browser client, which performs both an atomic update as well as a search of the repository for entities of matching type that contain similar strings. When the repository entries are not sufficiently informative to disambiguate among competing entities, the tool can bring up a general purpose document browser to display the section of a document from which any particular repository entry was derived. This information is maintained in each atomic entry in the repository.

What the user sees of the repository is a local client, integrated with the document-level annotation client. When the result of an annotation is to be added to the entity repository, the repository client performs both an atomic update as well as a search of the repository for similarly typed entities that contain similar strings. This supports the next activity, namely determining which of the existing entries may be coreferential (or otherwise related) with this newly annotated data.

It will often happen, especially with entities with few mentions and/or few "features" that the human annotator will need to refer to the original documents in which one or another of the repository entities were derived. All of the results from a query maintain their links back to the source documents from which they were derived, so the user may click on the appropriate returned entity to view the references to this entity in the context of the source document. In this case we are simply using a World Wide Web browser to retrieve and display these original source documents (shown in the upper right of the Figure 7).

### 5. Evaluation

There are three distinct ways in which one might want to evaluate cross-document information extraction. The different techniques stress different concepts, and are abstract. There are also parameters on the coreference evaluation task which must be set for instantiation (including a description of markables and categories into which markables will be sorted), but those will not be discussed here.

**Equivalence class identity**, as exemplified in the MUC6 coreference scoring procedure (Vilain et al., 1995), measures the extent to which an equivalence classification of spans of text matches a reference partitioning. The unit of scoring is the member of one of the classes, and the algorithm used to calculate the score calculates the minimum number of operations that join or split equivalence classes required to transform the hypothesis into the reference.

**Link coreference**, as exemplified in the B-CUBED scoring procedure (Bagga and Baldwin, 1998), compares *all* of the pair-wise links that are implied by a coreference annotation to those in the reference. A set of *n* entities can have up to  $\binom{n}{(n,2)}$  links, and a missing entity has an impact roughly proportional to the number of links that it removes from the set. A similar calculation is performed to determine this metric, in which the number of inconsistent links between a reference and hypothesis is determined.

**Entity description**, as exemplified in the MUC6 mapping-based template element and scenario template dynamic programming scoring procedures (Chincor and Dungca, 1995), attempts to determine what portion of the relevant set of *facts* about an entity have been accumulated. Each classification of entities produces a set of facts about each class. A (reference,hypothesis) pair of those sets is then used to acquire precision and recall scores using the obvious method. This is perhaps the simplest embedded form of the coreference task, and it directly measures the motivating goal of the cross document information extraction. In this case the *facts* being evaluated would be derived from the reference and hypothesis versions of the cross-document entity/relation database.

The infrastructure defined in our annotation scheme supports all of these evaluation techniques, and we plan to evaluate our system (as well as compute inter-annotator agreement scores) using all of them.

## 6. Using the Repository Server for Automated Processing

So far we have viewed this task from the point of view of manual annotation. However, the same entity repository and its network services can be used by automated natural language processing systems to perform cross-document information extraction. We have done this for our multilingual information extraction system, Alembic (Aberdeen et al., 1995), which has been provided with bindings to the various repository services that allow it to update, query, merge and split entities in the repository. This provides an opportunity for integrating the document-level information extraction component and the cross-document reasoning component, while keeping them modular and separable. In particular we wish to enable both the integration of different components into the document-processing pipeline, and also support the component-level evaluation of crossdocument information tracking.

The availability of the cross-document repository and its client query engine presents interesting opportunities for studying the way in which humans use these tools to resolve ambiguous coreferential relationships. We would like to see whether the techniques used by skilled human operators can be captured and modified for use by our automatic processing systems. Furthermore, there is an exciting opportunity for close human-computer interaction in support of cross-document information extraction. The repository is a shared resource between human *annotator* and information extraction *system*, and thus is a communication channel of sorts. The result of the dialog that takes place along that channel is a better representation of the targetted information available in a set of documents.

### 7. Conclusions

We have discussed the issues involved in crossdocument annotation and proposed a framework for addressing them. This framework has been implemented in a set of modular tools, which can be incorporated both in various cross-document annotation tasks, and also as an aid in the construction of end-to-end cross-document information processing systems. We hope that the cross-document annotation tool-set described here will be a valuable resource for annotating a wide range of corpus-wide linguistic phenomena.

### 8. References

- 1995. Sixth Message Understanding Conference (MUC-6). Defense Advanced Research Projects Agency, Columbia, Maryland: Morgan Kaufmann.
- Aberdeen, John, John Burger, David Day, Lynette Hirschman, Patricia Robinson, and Marc Vilain, 1995. Description of the alembic system used for MUC-6. In Proceedings of the Sixth Message Understanding Conference (MUC-6). Columbia, Maryland.
- Bagga, Amit and Breck Baldwin, 1998. Algorithms for scoring coreference chains. In Proceedings of the Linguistic Coreference Workshop at The First International Conference on Language Resources and Evaluation (LREC'98).
- Bird, Steven, David Day, John Garofolo, John Henderson, Chris Laprun, and Mark Liberman, 2000. AT-LAS: A flexible and extensible architecture for linguistic annotation. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC2000) (to appear)*. Athens, Greece: ELRA. To appear, this volume.
- Chincor, Nancy and Gary Dungca, 1995. Four scorers and seven years ago: The scoring method for MUC-6. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*. Columbia, Maryland.
- Day, David, John Aberdeen, Lynette Hirschman, Robyn Kozierok, Patricia Robinson, and Marc Vilain, 1997. Mixed-initiative development of language processing systems. In *Fifth Conference on Applied Natural Language Processing*. Washington, D.C.: Association for Computational Linguistics.
- Vilain, Marc, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman, 1995. A model-theoretic coreference scoring scheme. In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*. Columbia, Maryland.