A Corpus for Cross-Document Co-Reference

David Day,[1] Janet Hitzeman,[1] Michael Wick,[2]
Keith Crouch,[1] Massimo Poesio[3]
[1]The MITRE Corporation, Bedford, MA, USA
[2]University of Massachusetts, Amherst, MA, USA
[3]University of Essex, UK and University of Trento, Italy
E-mail: day@mitre.org, hitz@mitre.org, mwick@cs.umass.edu,
kcrouch@mitre.org, poesio@disi.unitn.it

Abstract

This paper describes a newly created text corpus of news articles that has been annotated for cross-document co-reference. Being able to robustly resolve references to entities across document boundaries will provide a useful capability for a variety of tasks, ranging from practical information retrieval applications to challenging research in information extraction and natural language understanding. This annotated corpus is intended to encourage the development of systems that can more accurately address this problem. A manual annotation tool was developed that allowed the complete corpus to be searched for likely co-referring entity mentions. This corpus of 257K words links mentions of co-referent people, locations and organizations (subject to some additional constraints). Each of the documents had already been annotated for within-document coreference by the LDC as part of the ACE series of evaluations. The annotation process was bootstrapped with a string-matching-based linking procedure, and we report on some of initial experimentation with the data. The cross-document linking information will be made publicly available.

1. Introduction

There has long been a need for a rich corpus that reflects both within-document co-reference as well as cross-document co-reference. While within-doc co-reference will indicate whether there are multiple references to the same entity within a document, cross-document co-reference will indicate whether two different documents contain references to the same entity. It remains an open research question as to the extent to which algorithms developed for within-document coreference resolution can be successfully applied to the cross-document case. This corpus should be able to contribute to answering this question, since it incorporates high quality within-document coreference annotations, as well as the links resolving entities across documents. Being able to accurately resolve entity mentions across diverse documents will prove useful in a variety of application areas. For example, the results of search engine queries containing names of individuals could be organized by the unique entities to which they refer, accelerating the user's ability to navigate to the desired documents

This corpus was created as part of the activities of the workshop entitled "Exploiting Lexical and Encyclopedic Resources for Entity Disambiguation." Hosted by the Johns Hopkins University Center for Language and Speech Processing during the Summer of 2007, and it will be made available to the public. The corpus was based on the complete English ACE2005 Entity Detection and Recognition (EDR) data set, available from the Linguistic Data Consortium (LDC) data sets (catalog numbers LDC2005E18 and LDC2005T06) and described in (Doddington, et al, 2004). The LDC created the annotations that capture the within-document co-reference chains for seven types of entities (persons, organizations, locations, geo-political entities, weapons and vehicles) and their three types of mentions.

2. Annotation Methods

In order to create the cross-document co-reference corpus, we made use of the previously developed Callisto/EDNA annotation tool. This is a specialized annotation task plug-in for the Callisto corpus annotation tool (http://callisto.mitre.org/). This Callisto client plug-in makes use of a web server (in our case, we used Tomcat) and Lucene web services plugins created for this task. The ACE2005 source texts and standoff annotations (in the ACE APF XML format) are hosted on a server and indexed using a customized Lucene document parser. The result of this process is that search engine clients can search the ACE2005 repository using structure-dependent queries, such as searching for strings within entity name mentions, and within entities of a particular type and sub-type. The Callisto/EDNA annotation tool provides an integrated interface where EDR-annotated documents can be examined and individual entities can be linked to other entities in the corpus.

Figure 1 shows the Callisto/EDNA graphical user interface. The tagged text is presented in the upper left hand corner to provide the annotator with context for the mention(s) to be linked. In the pane at the bottom of the interface are presented tables of all of the within-document ACE annotations already present in the document. Usually the entity table will be displayed here, enabling the annotator to sort and step through the entity-level annotations to be linked across the corpus. The ACE entity annotations are color-coded: black indicates that the phrase is not a candidate for co-reference, green indicates that the term

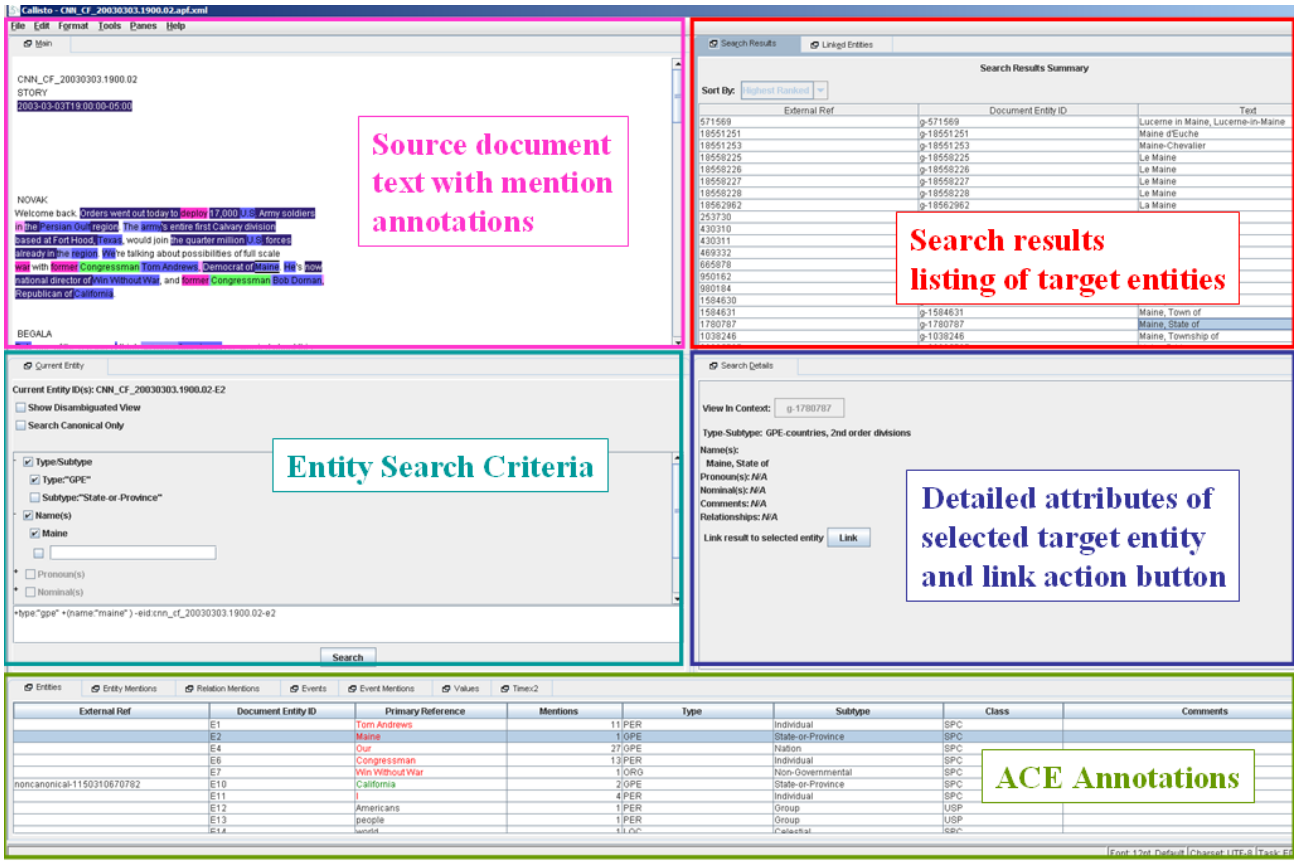

Figure 1: The Callisto/EDNA annotation graphical user interface

has already been co-referenced, and red indicates that the term has yet to be co-referenced. Clicking on an ACE annotation entry fills in the Entity Search Criteria with whatever is known about that ACE entity, e.g., the type and subtype. Clicking on the SEARCH button results in a list of search results that are potential target entities for co-reference across the complete document collection, which are then displayed in the upper right hand side search results pane. Clicking on a target entity brings up a set of detailed attributes for the annotator to consider about the selected target entity, and a LINK button that enables the annotator to co-reference the ACE annotation with the target entity, if desired. Figure 2 shows a close-up of the search query pane, which can be further manipulated by the user to broaden the search to alternate spellings, alternate entity types, etc. Figure 3 illustrates how the target entity details are displayed.

There has been some exploratory cross-document annotation of ACE documents in the past. The results of these experiments were that some types of entities posed particularly difficult for annotators to resolve, especially those for which no name was used as a mention anywhere in the document. To make this annotation process tractable, we configured the Callisto/EDNA tool to focus the annotation process on entities that met the following criteria:

(1) The entity had at least one mention of type NAME with a document;
(2) The entity was of type PER, ORG, GPE or LOC.

To expedite the annotation process, we decided to apply an initial automated pre-annotation (cross-document linking) process prior to manual annotation. We had observed in early efforts that much of the time invested by the human annotator was in physically linking frequently occurring entities to each of the numerous entities in other documents where such entities were mentioned. For example, the President of the United States occurred in a significant percentage of the ACE documents, and the annotation of this phrase would necessitate a laborious process of stepping through the physical clicking (actually a whole cascade of user mouse actions) of many highly probable co-referring entities. The automatic pre-processing procedure was written in Java to load the complete ACE2005 corpus EDR annotations into memory, after which it proceeded to link each pairwise entity just in case those two entities were of exactly the same TYPE and SUB-TYPE and the entities shared at least one mention of type NAME whose strings were identical (using a case-sensitive string comparison test). We were concerned that this procedure would produce inappropriate links, but in actuality it produced very few. The biggest error it made was to link together all

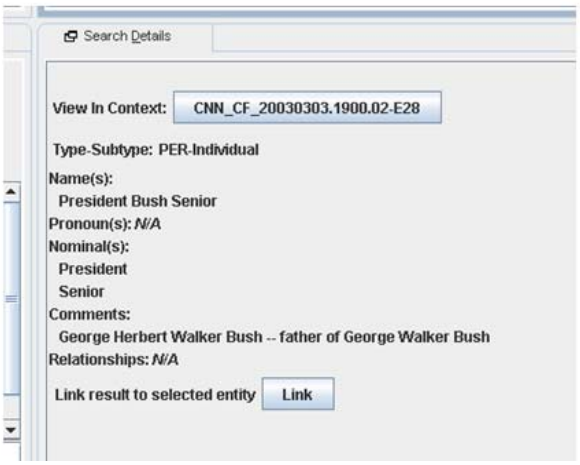

Figure 2: The search query pane

phrases "Anonymous speaker," but this was in part caused by the annotator of within-document co-reference incorrectly linking them as well.

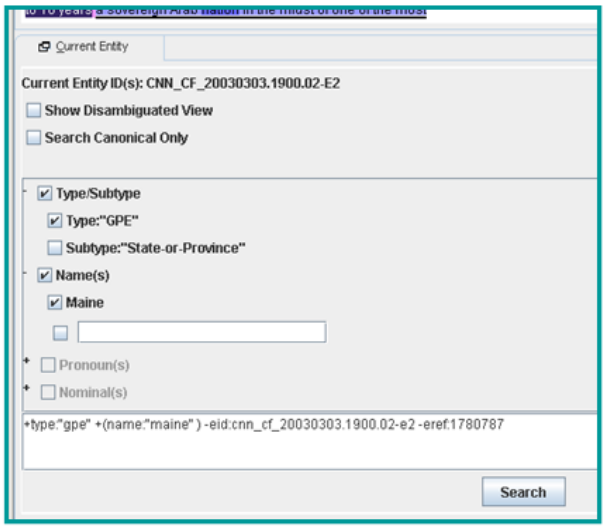

Figure 3: The entity details pane

The Callisto/EDNA annotation tool provided the ability for the annotator to quickly review and, if warranted, remove any links made earlier, whether by this automatic process or by a human annotator. As a bonus, annotation became faster after this pre-processing step, as links could be made at the group level, where a group would be all those entities sharing a common name, type and subtype. For example, a single linking action between an entity cluster with the named entity Bush and another entity cluster with the named entity the "W" would enable all of the associated entities in various documents to be linked.

3. The Corpus

The resulting corpus (derived from the ACE2005 English EDT corpus) consists of approximately 1.5 million characters, 257,000 words and 18,000 distinct document-level entities (prior to cross-document linking), and approximately 55,000 entity mentions. The document-level entities are distributed across entity types in the following way: Person 9.7K, Organization 3K, Geo-Political Entity (GPE) 3K, Facility 1K, Location 897, Weapon 579, Vehicle 571. The entity *mentions* are distributed across mention type in this way: Pronoun 20K, Name 18K, Nominal 17K. Those entities that satisfy the constraints required for them to be included in the cross-document annotation process number 7,129. After the combination of automatic and manual annotation, the number of cross-document entities numbers 3,660. Of these, 2,390 are entities that are mentioned in only one document. Only entities of type person, location, geo-political entity (GPE) and organization were included in the cross-document linking captured in this corpus, and only those for which at least one name mention was present within a document.

4. Previous Work

To our knowledge the first effort for creating corpora to train and evaluate cross-document co-reference resolution algorithms was that initiated by Breck Baldwin and Amit Bagga (Bagga, et al, 1998), in which they created the so-called John Smith corpus. Bagga and Baldwin used the relatively common name John Smith to find documents that were about different individuals in the news. The two main benefits of the present work are (1) by performing the cross-document co-reference annotation on top of the already fully-annotated ACE2005 corpus, we have high-quality co-reference information at both the within-doc and cross-document levels; (2) the size of this corpus is significantly larger than the previously available data sets.

There have been other data sets created that address the general problem of identifying entities across documents, but these have involved annotation schemes that do not capture all of the named, nominal and pronominal mentions of the entities. For example, in the recent SemEval "Web People" evaluation (Artiles, et al, 2005), disambiguation is annotated at the level of documents that are (mostly) "about" a particular person, although some systems exploited mention-level features and local contexts (e.g., Heyl & Neumann, 2007). The annotation schemes adopted to support author identification in citations (Lawrence, et al, 1999) is at a similar mention level, but the texts themselves, collections of bibliographic citations, are quite specialized.

5. Data Format

The Callisto/EDNA tool was designed to output files in a format similar to the ACE APF format. An *external_link* element is added within the entity element of the APF data, which associates a particular document-level entity annotation with an identifier. This identifier (the EID attribute) defines an equivalence class of entities and their document-level mentions, all of which refer to the same entity. An example is shown in Figure 4.

```
<entity CLASS="SPC"
   ID="AFP_ENG_20030323.0020-E62"
   SUBTYPE="Individual" TYPE="PER">
 <entity_mention
   ID="AFP_ENG_20030323.0020-E62-86"
   LDCTYPE="NAMPRE" TYPE="NAM">
  <extent><charseq END="3161"
    START="3152">John Wayne</charseq> ...
 <external_link EID="1772"
                RESOURCE="elerfed-ed-v1"/>
 </entity>
```

Figure 4. Example APF Cross-document linking annotation

6. Observations

One side-effect of performing cross-document co-reference was that errors in the within-document co-reference became apparent. Examples include the co-referencing of "Scott Peterson" and "Laci Peterson," probably because they were each referred to as "Peterson" on some occasion.

One desirable side effect of this annotation is that it clusters proper names with their corresponding nicknames. For example, "Bama" for "Alabama", and "Qland" or simply "Q" for "Queensland". These nicknames will not be found in a gazetteer, but the cross-document corpus will give a system the means to find the full name, which will give it the ability to map a location using a gazetteer.

7. Assessing Ambiguity

One might expect that references to entities such as George Bush would be ambiguous between George Sr. and George Jr. However, because we focused on a small period of time, all mentions of George Bush refer to the same underlying entity. There are also a large number of entities that occur infrequently, often represented by just a single mention in the entire corpus, such as Wayne Gretzky.

To test the ambiguity of the dataset, we implemented a discriminatively trained clustering algorithm similar to the one described by Culotta et al (2007). We measured the cross-document co-reference performance on the gold standard intra-document co-reference chains from the reserved test set of documents. Our features included neighboring words (and mentions) in local context windows around each chain, as well as words and mentions from the documents. Additionally, features compare the canonical mentions in each chain, testing for substring matches. We were able to achieve F1 performance as high as: 0.96 (BCubed), 0.91 (Pairwise) and 0.89 (MUC).

This performance is relatively high, indicating a low ambiguity level. However, this performance is the result of clustering the gold-standard within-document co-reference chains; given predicted chains this problem becomes much more challenging. Additionally, ambiguity can be retroactively injected into the dataset as desired. Methods analogous to pseudo-words in the word sense disambiguation task have been applied to co-reference to introduce ambiguity while still retaining many statistical properties of the original dataset (Mann and Yarowsky 2007).

8. References

J. Artiles, J. Gonzalo, F. Verdejo (2005). A Testbed for People Searching Strategies in the WWW. SIGIR 2005 Conference. Special Interest Group on Information Retrieval.

Bagga, A., Baldwin, B. (1998). Algorithms for Scoring Co-Reference Chains. In *Proceedings of the Linguistic Coreference Workshop at The First International Conference on Language Resources and Evaluation (LREC'98),* pp. 563-566.

Culotta, A., Wick, M., Hall, R., McCallum, A. (2007). First-order Probabilistic Models for Coreference Resolution. *Proceedings of Human Language Technology (HLT).*

Heyl, A. and Neumann, G. (2007). "DFKI2: An Information Extraction Based Approach to People Disambiguation." Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), pages 137—140, Prague.

Doddington, G. (2001). Value-based Evaluation of EDT. *Technical Report of the ACE 6-Month Meeting.*

Lawrence, S., Giles, C. L., & Bollacker, K. (1999). *Digital libraries and autonomous citation indexing.* IEEE Computer, 32, 67–71.

Mann, G., Yarowsky, D. (2007). *Unsupervised Personal Name Disambiguation.* CoNLL, 2003, Edmonton, Alberta, Canada.

Solomonoff, A., Mielke, M., Schmidt, M., Gish, H. (1998). Clustering speakers by their voices, *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 757–760.