Linguistic Resources and Evaluation Techniques for Evaluation of Cross-Document Automatic Content Extraction

Stephanie Strassel[1], Mark Przybocki[2], Kay Peterson[2], Zhiyi Song[1], Kazuaki Maeda[1]

(1) Linguistic Data Consortium (LDC)
3600 Market Street, Suite 810
Philadelphia, PA 19104 USA
{strassel, zhiyi, maeda}@ldc.upenn.edu

(2) National Institute of Standards and Technology (NIST)
Information Access Division, Speech Group
100 Bureau Drive, Stop 8940
Gaithersburg, MD 20899-8940 USA
{mark.przybocki, kay.peterson}@nist.gov

Abstract

The NIST Automatic Content Extraction (ACE) Evaluation expands its focus in 2008 to encompass the challenge of cross-document and cross-language global integration and reconciliation of information. While past ACE evaluations were limited to local (within-document) detection and disambiguation of entities, relations and events, the current evaluation adds global (cross-document and cross-language) entity disambiguation tasks for Arabic and English. This paper presents the 2008 ACE XDoc evaluation task and associated infrastructure. We describe the creation of development and test data to support the evaluation, focusing on new approaches required in data selection, annotation task definition and annotation software; and we conclude with a discussion of the metrics developed to support the evaluation.

1. Introduction

The objective of the NIST Automatic Content Extraction (ACE) series of evaluations is to develop human language understanding technology that provides automatic detection and recognition of key information about real-world entities, relations, and events that are mentioned in source data. To date, the focus of identification of these real-world objects has been limited to being linked to the document in which they are mentioned, while coreferencing the mentions of the same real-world object across documents has been regarded as a not-yet-reachable goal. This focus is changing in 2008 with the ACE evaluation scaling up to cross-document and cross-language global integration and reconciliation of information.

To this end, the 2008 ACE evaluation will not only include entity and relation detection and disambiguation for Arabic and English, but also cross-document and cross-language global integration and reconciliation of information. For cross-document and cross-language entity disambiguation tasks (hereafter, XDoc), system output will be evaluated only for person (PER) and organization (ORG), and only for documents in which the "target entities" are mentioned by name. Target entities refer to a carefully selected list of pre-defined entities of interest.

To support this new challenge there is a need for new data, annotation tools, guidelines and metrics.

2. Data

Assembling appropriate data sets for cross-document entity detection requires a careful and targeted process. The appropriateness of the data sets is determined by the intent of the evaluation. In early ACE evaluations, training and test documents were selected using a stratified random sample of material from the available data sources and epochs.

Starting with ACE 2005, new evaluation requirements and a growing inventory of targeted annotation types and subtypes motivated a more careful data selection strategy in order to provide a minimum density of occurrences of each type in the corpus. Accordingly, LDC developed a series of machine and human algorithms to perform targeted data selection. The ACE XDoc task presents an even greater challenge for document selection, since we must consider not only the density of entity types in the corpus as a whole, but also the frequency of specific entities, both within and across documents. In fact, data selection is emerging as the most significant annotation challenge for the ACE 2008 XDoc task.

The ability to process large amounts of data is critical for ACE systems. This capability is especially important for disambiguation across multiple documents and languages. Whereas previous ACE evaluations tested system performance on a couple hundred carefully selected documents per language, the 2008 ACE evaluation corpus will be on the order of 10,000 documents per language. The corpus will be drawn from a much larger data pool comprised of multiple genres including newswire, weblogs, Usenet newsgroups and bulletin boards, and transcripts of broadcast news, talk shows and conversational telephone speech. The large and diverse data pool is expected to capture a greater range of entity mention variations (including alternative name forms, aliases, transliterations, etc.).

While ACE systems are expected to extract data on all ACE entities and relations for all documents, scoring is done only on a select subset of the occurrences. The selective scoring will be driven by pre-evaluation selection of a target list of 50 entities per language. To support the goals of the evaluation, the target entity list must possess the following features (among others):

• each entity is mentioned between 5 and 100 times in the 10,000 word corpus
• some entities appear in both the English and Arabic
• some entities have aliases that occur in the corpus (for instance, Ilich Ramírez Sánchez might be selected if it is known that Carlos the Jackal is also mentioned in the corpus)
• some entities have orthographic variation in the corpus (for instance, Mu'ammar Al-Qadhafi might be selected if it is known that Muammar al-Gaddafi also occurs)
• some entities should be confusable with other distinct entities (for instance, Michael Jordan might be selected if it is known that the corpus contains mentions of "Michael Jordan the US basketball player" and mentions of "Michael Jordan the English football player"

The complete data selection task for ACE 2008 entails identification of 50 target entities that possess the desired properties, and selection of the 10,000 documents that provide maximal coverage of the target entities. The selection task is something of a Catch-22: we need to know which entities are targeted so that the best set of documents can be selected, and we must simultaneously ensure that the entities are actually mentioned within the chosen documents. In response, LDC has defined a two-part data selection strategy. During the Unstructured Exploration phase, annotators use world knowledge, information about name frequencies in existing LDC annotated corpora, and simple automated techniques (for instance, histograms of arbitrarily-long capitalized strings, sorted for similarity) to generate a very large list of potential entities that are likely to occur in the data pool. For each entity, annotators create an entity profile that contains basic information such as known aliases, orthographic variants and key facts about the entity. The entity profile also identifies an entity handle that serves as the standard name reference for this entity for all subsequent tasks. Entity profiles may be supplemented with information drawn from outside the data pool including knowledge sources like Wikipedia. In the Structured Exploration phase, annotators work through the list of candidate entities, issuing queries against the assembled data pool to determine: a) whether a given entity is mentioned in the corpus, b) what is the frequency of mentions for that entity; and c) what are the name properties for that entity (e.g. is there an alias in the data pool?).

LDC has designed the CEToolkit, a customized corpus exploration and annotation software package, to facilitate searching. Annotators issue queries across the indexed data pool, and CEToolkit's embedded search engine returns a list of relevance ranked documents. Word, phrase, example-based searching as well as boolean search syntax are all supported, and searches may be further restricted by date, genre, and/or source. Search terms are highlighted in the resulting documents to facilitate subsequent annotation. When a document of interest is found, the annotator makes a series of judgments about the entity properties observed in the document, along with a value rating (1-5 stars) for the document. The CEToolkit also allows users to capture and log arbitrary strings of text from the document, which can then be flagged as “potential new entity” or “related to current entity”, providing additional information for future corpus exploration. All annotator judgments are logged to a MySQL database. The CEToolkit supports remote users, allowing corpus exploration to be distributed across multiple sites.

At the conclusion of corpus exploration, LDC will use a semi-automated process to analyze the resulting database and select the 250 names for each language that best illustrate the desired range of entity properties. The next stage is data pruning. LDC will use the list of 250 target entities to automatically select the best 10,000 documents from the much larger data pool. The pruning algorithm will address each requirement -- frequency of entity mentions, orthographic variation, etc. -- selecting documents that have the highest value given the 250 target entities. Various weighting parameters will be set; for instance, documents that contain mentions of multiple target entities will be weighted more heavily than documents that contain only one entity. Because confusable entities (distinct entities with the same orthographic name) are likely to be very rare in the data pool, the algorithm will heavily weight document clusters that contain mentions of confusable entities. The resulting 10,000 document set will constitute the evaluation corpus for the ACE 2008 XDoc pilot task. The list of 250 entities will also be subsampled to identify the 50 entities that constitute the target entity list for the current evaluation.

3. Annotation

Manual XDoc annotation for the full 10,000 document corpus, across all targeted entities, would be prohibitively expensive. Instead, LDC will perform full entity and relation annotation (LDC 2008a, LDC 2008b) plus full manual cross-document coreference on a 400-document subset of the full evaluation corpus. These documents will be automatically sub-sampled from the evaluation corpus using a variety of methods, including a refined implementation of the data pruning technique described above.

Even on a constrained document set the manual XDoc task presents a major cognitive challenge for annotators, complicated still further by the fact that entities in these documents are specifically targeted to exhibit a wide variety of surface representations. XDoc entity

coreference will be facilitated by use of the Entity Disambiguation and Normalization Annotation (EDNA) component of the MITRE Callisto toolkit (Day et. al. 2006). EDNA takes ACE entity-annotated documents as input. The tool allows users to easily search the repository of annotated documents for other probable mentions of a given entity, and to quickly merge mentions of the same entity across documents. The resulting annotation of the 400 document subset will serve as the basis for XDoc system evaluation. While manual XDoc annotation is likely to result in high precision (accuracy), it will undoubtedly suffer in terms of recall. Because manual XDoc annotation cannot be exhaustive, the evaluation will also include a manual post-hoc adjudication component.

4. Metrics

The last ACE evaluation to incorporate a cross-document task occurred in 2002. The test condition was limited to type PERSON and required that a subset of entities and relations be uniquely identified independently of the document that mentioned them. Scoring was simplified by requiring systems to use specific entity IDs that were provided in a seed database. The lone evaluation metric was the ACE_VALUE formula.

In the current ACE effort, there is a renewed focus on alternative metrics. In addition to the ACE_VALUE formula, system performance will be evaluated using the B-CUBED algorithm combined with post-hoc adjudication of "system-identified" entities that were not previously annotated. For successful XDoc scoring, a required change for ACE systems will be to assign a single entity-ID (equivalence class ID) to each within-document entity record.

"ACE_VALUE" (Doddington et. al. 2004) is determined by the global optimum mapping of system output to manually annotated reference data, where each correctly identified attribute is credited with a specific value, according to the ACE value formula. Errors in the attributes of mapped system output fail to add to the possible value and spurious system output tend to provide negative value. A parameter set is defined for all ACE tokens and their corresponding attributes. ACE_VALUE has been the primary metric of all previous NIST ACE evaluations. Alternatives to the maximal mapping approach are being considered, including reducing mention attributes to a listing of documents that mention the targeted entity.

"B-CUBED" (Bagga & Baldwin 1998) is a set-based algorithm that relies on the intersection between reference and system sets where the measures is how well system outputs are clustered without the need for determining an explicit one-to-one mapping of system output to reference annotation. Each item in an equivalence set contributes a fractional amount as a function of the missing items. B-CUBED computes the precision and recall of the ACE objects defined in the reference annotation. The B-CUBED algorithm will be added to current ACE scoring software, (NIST 2008a) and it can be tailored to provide scoring based on document lists or entity mentions. In addition to the traditional precision and recall, the B-CUBED algorithm will be enhanced to report a value-weighted precision and recall, similar to that of the ACE_VALUE metric.

See the appendices of (NIST 2008a) for detailed formulas and descriptions of the ACE_VALUE and B-CUBED metrics.

"Post-Hoc ADJUDICATION" will assess the precision performance of a system by manually judging whether or not a system-identified entity is truly a mention of the target entity. While reference annotation for the approximately 50 targeted entities will be limited to the occurrences in 400 documents, systems will produce output for all 10,000 documents. The maximal mapping between the reference entities and the system entities will produce a list of "spurious" system entities for each of the target entities (presumably mentions that exist outside the 400 document set that is pre-annotated). Each spurious system mention will be manually evaluated "thumbs-up or thumbs-down" as to whether or not it is a valid mention. These decisions can be tallied and fed back into the ACE_VALUE and B-CUBED formulas.

Each of the three proposed ACE metrics will be tested using a mini-corpus designed to reflect issues to be encountered in the planned full ACE evaluation corpus.

Some time after the conclusion of the ACE08 evaluation, the resulting XDoc resources including data, annotations, LDC and NIST software, guidelines, task definitions and related evaluation infrastructure will be made available through the usual NIST and LDC distribution mechanisms.

5. References

Bagga, A. and Baldwin B. (1998). Entity-based Crossdocument Coreferencing Using the Vector Space Model. 17th International Conference on Computational Linguistics (CoLing-ACL). Montreal, Canada. 10-14 August, 1998, 79-85.

Day, D., McHenry, C., Kozierok, R., Riek, L. (2004). Callisto: A configurable annotation workbench. LREC 2004: Fourth International Conferenceon Language Resources and Evaluation

Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., Weischedel, R. (2004). Automatic Content Extraction (ACE) program - task definitions and performance measures LREC 2004: Fourth International Conference on Language Resources and Evaluation

Linguistic Data Consortium. (2008a). English Entity

Annotation Guidelines V6.1 http://projects.ldc.upenn.edu/ace/docs/English-Entities-Guidelines_v6.1.pdf

Linguistic Data Consortium. (2008b). English Relation Annotation Guidelines V6.0 http://projects.ldc.upenn.edu/ace/docs/English-Relations-Guidelines_v6.0.pdf

NIST (2008.) ACE08 Evaluation Software ace08-eval-v01. ftp://jaguar.ncsl.nist.gov/ace/resources/ace08-eval-v01

NIST (2008a), ACE 2008 Evaluation Specification Document http://www.nist.gov/speech/tests/ace/2008/doc/