Automatic Content Extraction 2008 Evaluation Plan (ACE08)

Assessment of Detection and Recognition of **Entities** and **Relations** Within and Across Documents

1. Introduction

The objective of the Automatic Content Extraction (ACE) series of evaluations has been to develop human language understanding technologies that provide automatic detection and recognition of key information about real-world entities, relations, and events in source language text, and to convert that information into a structured form, which can be used by follow-on processes, such as classification, filtering and selection, database update, relationship display, and many others.

An ACE system produces information about objects discussed in the source language text. The strings of text are not the objects, but are merely mentions of the real-world objects about which information should be extracted. These objects have included, over the course of the evaluations, various types of entities, relations, events, values, and temporal expressions. The emphasis has been on object co-reference resolution, such that all data pertaining to the same unique ACE object are collected into a single XML-formatted "record" on a per document basis. Information about the same object from multiple documents and across multiple languages is associated through a common object identifier (equivalence class). Section 2 of this plan defines the objects of interest for the ACE 2008 evaluations.

In brief, though, the 2008 ACE evaluation will involve within-document and cross-document tasks in Arabic and English. Within-document object detection and recognition will be scaled back to only entities (EDR) and relations (RDR), and will not include event, value, or timex2 objects. Only the original five ACE entity types (people, organizations, geo-political entities, facilities, and locations) will be addressed for within-document EDR, while cross-document EDR will be limited to only person (PER) and organization (ORG) entities, and only for those documents in which they are mentioned by name. The evaluation for within-document relations will remain the same, while cross-document RDR will be limited to only those relations that are between PER and ORG entities that are named in the documents. New to this year's evaluation is the request that systems give confidence values [0 to 1 likelihood] for entity and relation extractions.

Also of interest this year is the ability to process large amounts of data, especially for disambiguation across multiple documents. Therefore, the 2008 ACE evaluation corpus will be on the order of 10,000 documents per language. This size will allow for the occurrence of a greater variety of entity mentions (including alternative name forms, aliases, misspellings, and transliterations) and for more entities to occur in more documents. Evaluation will be performed over only a limited subset of documents selected from the total evaluation corpus. This subset of documents will be made as large as can be practically annotated. Results from these documents will be made available to the evaluation participants prior to the evaluation workshop for their study and analysis. Also, the submissions from all systems will be pooled and used in a post-evaluation assessment phase to help validate and refine the original reference annotation. The resulting refined answer keys will be made available to the participants prior to the evaluation workshop.

2. Task Definitions

The ACE08 tasks are split into two groups, according to whether the context is local (limited to the document being processed) or global (across documents). The former provides continuity with previous evaluations, while the latter adds new challenges for linking entity and relation information across separate documents within each language.

For 2008, the ACE object categories will be limited to entities and relations. Systems must extract information about these objects from language data in documents and then output that information in a structured form. For a complete description of the ACE objects and their attributes, refer to the ACE annotation guidelines [1] prepared by the Linguistic Data Consortium. Within-document detections are output in ACE Program Format (APF). The XML DTD for this format may be found on the NIST ACE web site.[2]

2.1 Local Entity Detection and Recognition

The Local Entity Detection and Recognition task (LEDR) requires that ACE entities mentioned in source language data be detected, and that selected information about these entities be recognized and merged into a consolidated XML representation on a per entity and a per document basis. The information comprises the attributes and the mentions of that entity. For the local EDR task, each document is processed separately and entities that are mentioned in different documents are treated as different entities (by assigning unique document-specific ID's to them), even if in the real world they are the same entity.

2.2 Entity Attributes

Entity attributes are currently limited to *type*, *subtype*, *class*[3], and the set of *name*(s) used to refer to the entity. Optionally (but preferably, a *confidence value* (confidence in the existence of the entity in the document will also be given. The allowable ACE entity classes are listed in Table 1. Entity types and subtypes are given in Table 2. Entities may have only one class, one type, and one subtype. These are described in detail in the annotation guidelines.

There are no limits on the use of inference or world knowledge in detecting and recognizing entities. However, there are restrictions against examining or training on the evaluation test data. Any extraction determination should represent the system's best judgment of the source author's intention.

It often happens that different entities may be referred to by text strings that appear to be the same name. However, such entities are

[1] http://www.ldc.upenn.edu/Projects/ACE/Annotation

[2] http://www.nist.gov/speech/tests/ace/2008/doc

[3] Only "specific" entities (class="SPC") are assigned a non-zero value during evaluation and therefore systems need output only SPC entities for evaluation. Correct recognition of an entity's class is important for good performance, though, because the value of the output will be reduced for each SPC entity that is incorrectly classified as a non-SPC entity, and vice versa.

regarded as separate and distinct for the purposes of the ACE evaluation. For example, in the sentence "*Miami is growing rapidly*", Miami is a mention of a geo-political entity (GPE) named "Miami", whereas in the sentence "*Miami defeated Atlanta 28 to 3*", Miami is a metonymic mention of a sports organization entity named "Miami Dolphins" and is distinct from the Miami GPE entity.

Table 1 ACE08 Entity Classes

Type	**Description**
SPC	A particular, specific and unique real world entity
GEN	A generic entity (i.e., a broad "class" of entity)
NEG	A negatively quantified (usually generic) entity
USP	An underspecified entity (e.g., modal/uncertain/…)

Table 2 ACE08 Entity Types and Subtypes

Type	**Subtypes**
FAC (Facility)	Airport, Building-Grounds, Path, Plant, Subarea-Facility
GPE (Geo-Political Entity[4])	Continent, County-or-District, GPE-Cluster, Nation, Population-Center, Special, State-or-Province
LOC (Location)	Address, Boundary, Celestial, Land-Region-Natural, Region-General, Region-International, Water-Body
ORG (Organization)	Commercial, Educational, Entertainment, Government, Media, Medical-Science, Non-Governmental, Religious, Sports
PER (Person)	Group, Indeterminate, Individual

2.3 Entity Mentions

The requirement for outputting entity mentions is conditioned upon the task. For GEDR (see below) entity mention output is not required. For LEDR, entity mention output ***is*** required.

All mentions of each ACE entity are to be detected and output along with the entity attributes. The types of these mentions are listed in Table 3. The output for each entity mention includes the mention *type*, its *extent*, the location of a *head* within the extent, and optionally the mention *role* and *style*. Mention *style* is either literal or metonymic. This is currently encoded in the ACE Program Format (APF) as an attribute called "metonymy mention", which is either true (for metonymic style of reference) or false (for literal style of reference). The default style is literal. Mention attributes and their possible values are described in detail in the annotation guidelines.

Table 3 ACE Mention Levels (Categories)

Type	**Description**
NAM (Name)	A proper name reference to the entity
NOM (Nominal)	A common noun reference to the entity, or a phrasal description of the entity
PRO (Pronominal)	A pronominal reference to the entity

2.4 Diagnostic LEDR

In order to assist in assessing the quality of the co-reference resolution components of LEDR processes, participants in the LEDR task will be encouraged to run their system for a follow-on diagnostic task. In this task, ground-truth entity mentions will be provided for the systems to co-reference. This ground-truth data will only be provided to sites participating in the LEDR task, and only after the submission of the results of their LEDR processing.

2.5 Entity Mention Detection (EMD)

LEDR systems will also be scored for Entity Mention Detection accuracy. This evaluation will assess a system's ability to detect isolated mentions of ACE-defined entities in the source language and to recognize and output selected attributes and information about these entity mentions. This data includes the entity type, subtype, and class, as well as mention level (NAM, NOM, PRO) and beginning and ending offsets of the mention in the document.

In this task, each entity mention is treated independently, and, therefore, is given a unique entity identifier. Nevertheless, co-reference still remains an important issue because each entity mention must be a mention of an entity within the set of ACE entities. Section 2.3 describes entity mentions. Table 3 lists the mention levels (categories).

2.6 Global Entity Detection and Recognition

The global entity detection and recognition task (GEDR) requires cross-document coreference resolution of entities of type PER and ORG, based on name-level references. The name-level references can include long and short forms of the name, variant spellings, misspellings, transliterations, aliases, and nicknames. These should be output in the entity name attribute XML element of the ACE Program Format. Output of entity mentions is not required for GEDR. Global reconciliation is accomplished by using the same unique global (corpus-wide) entity identifier as the entity ID attribute for every document in which the same entity is mentioned. Refer to appendix C for a condensed version of the apf dtd used for system output.

Note, however, that, for scoring purposes, a metonymic NAM mention[5] is not a proper name for an entity, and, therefore, is not reflected in the entity name attribute XML element for that entity. For instance, "Washington" is not a proper name for the "United States". **Also, note** that entities with no literal NAM mentions (i.e., those with only metonymic NAM mentions) are ***not*** NAM level entities, and thus are excluded from scoring in GEDR.

[4] Geo-Political Entities deserve a little explanation and historical background. Originally, GPE's were not part of the ACE entity inventory. However, during the initial annotation exercises, it became clear that the same word would often imply different entity types – sometimes *location* (as in "the riots in Miami"), sometimes *organization* (as in "Miami imposed a curfew"), sometimes as *person* (as in "Miami railed against the curfew"). Even more troublesome, co-reference was sometimes observed between different underlying entity types (as in "Miami imposed a curfew because of its riots"). These issues gave rise to the definition of the hybrid Geo-Political entity type. This type can be viewed as somewhat synthetic and ad hoc, but there is also support for its conceptual reality, for example by the use of co-reference in joining different entity types.

[5] English Annotation Guidelines for Entities, Chapter 6. Marked as TYPE="NAM" METONYMY_MENTION="TRUE".

2.7 Local Relation Detection and Recognition

The Local Relation Detection and Recognition task (LRDR) requires that ACE relations that are mentioned in the source language data be detected, and that selected information about these relations be recognized and merged into a unified XML representation for each detected relation. Note, however, that for ACE08 no time data will be required for relation extraction Please refer to the annotation guidelines for detailed information about determining ACE relations.

An ACE relation is a relationship between two ACE entities, which comprise the main "arguments" of the relation. Some relations are symmetric, meaning that the ordering of the two entities does not matter (e.g., "partner"). However, others are asymmetric, so the order of the arguments does matter (e.g., "subsidiary"). For these relations, the entity arguments must be assigned to the correct argument role (Arg-1 or Arg-2).

The information that an ACE system must output for each relation is specified in the relation attributes, arguments, and mentions (see the following three sections for details). For local RDR, relations that are mentioned in different documents are presumed to be different relations. Therefore, information extracted from a specific document must be assigned to a document-specific relation; i.e., a relation with a document-specific ID that uniquely determines the document and the relation. The relation arguments must also be document-specific objects (entities).

2.8 Relation Attributes

Relation attributes are the relation *type*, *subtype*, *modality*, and *tense*. The ACE relation types and subtypes are listed in Table 4. Relations may have only one type and one subtype.

Table 4 ACE08 Relation Types and Subtypes
(Relations marked with an * are symmetric relations.)

Type	Subtype
ART (artifact)	User-Owner-Inventor-Manufacturer
GEN-AFF (General affiliation)	Citizen-Resident-Religion-Ethnicity, Org-Location
METONYMY*	*None*
ORG-AFF (Org-affiliation)	Employment, Founder, Ownership, Student-Alum, Sports-Affiliation, Investor-Shareholder, Membership
PART-WHOLE (part-to-whole)	Artifact, Geographical, Subsidiary
PER-SOC* (person-social)	Business, Family, Lasting-Personal
PHYS* (physical)	Located, Near

2.9 Relation Arguments

Relation arguments are identified by a unique ID and a role. The roles of the two entities being related are "Arg-1" and "Arg-2". The correct assignment of these roles to their respective arguments is important, except for symmetric relations (which are identified in Table 4 by an asterisk). There may be only one Arg-1 and one Arg-2 entity. The list of allowable argument roles for relations is given in Table 5.

Table 5 Argument roles allowable for relations

Allowable Relation Roles	
Arg-1	Arg-2
Time Mention (not used in ACE08)	

2.10 Relation Mentions

A relation mention is a sentence or phrase that expresses the relation. The extent of the relation mention is defined to be the sentence or phrase within which the relation is mentioned. A relation mention must contain mentions of both of the entities being related. Although recognition of relation mentions is not evaluated directly, it is one of the ways that system output relations are allowed to map to reference relations. Thus, correct recognition of relation mentions is potentially helpful in evaluation.

2.11 diagnostic LRDR

In order to assist in assessing the quality of the co-reference resolution components of LRDR processes, participants in the LRDR task will be encouraged to run their system for a follow-on diagnostic task. In this task, ground-truth entities will be provided, which the systems can use for finding relevant relations. This ground-truth data will only be provided to participants in the LRDR task, and only after a site has submitted its LRDR results

2.12 Relation Mention Detection

All LRDR systems will subsequently be scored for Relation Mention Detection (RMD) accuracy. RMD requires systems to find independent mentions of ACE relations, and to output their attributes and arguments. Each mention of an ACE relation is treated independently, and, therefore, is given a unique relation identifier. Section 2.10 describes relation mentions.

2.13 Time Stamping of Relations

ACE08 will not include a separate evaluation of timex2 performance, and evaluation of relations will ignore timex2 arguments, if they are included in relation output.

2.14 Global Relation Detection and Recognition

The global relation detection and recognition task (GRDR) requires that the same unique ACE relation be found across documents for the same globally reconciled entities (limited to GEDR entities only – PER and ORG in documents where they are named). A unique relation is defined by the relation type, subtype, and a pair of entity arguments. The REFID of a global entity must be used as the relation argument for at least one of the relation arguments. Global reconciliation is accomplished by using the same unique global (corpus-wide) relation identifier as the relation ID attribute for every document in which the same relation is mentioned. Output of relation mentions is not required for GRDR. Refer to appendix C for a condensed version of the APF DTD used for system output.

3. Corpus Support

Annotated source language data is being provided to support research and evaluation. This includes training corpora (development test set) and an evaluation test corpus. ACE corpora are assembled from a variety of sources, including radio and TV broadcast news, talk shows, newswire articles, internet news groups, web logs, and conversational telephone speech.

3.1 The ACE 2008 Training Corpus

For the local detection and recognition tasks, ACE08 will use the same training data as was used for ACE07, except that the

languages are restricted to Arabic and English, and the tasks are restricted to entities and relations. Annotations for times, values, and events are not relevant for ACE08, but will likely be applicable again in future evaluations. For the global tasks, a special subset corpus is provided as an example of what is desired.

The Linguistic Data Consortium provides annotated training data[6] for ACE system development. The data is taken from a variety of sources and is available for tasks in Arabic and English. See Table 6 for the training corpus statistics.

The ACE training and evaluation data was selected using a targeted process. Rather than choosing files at random for annotation, as was done in some past ACE evaluations, this year's tasks required annotation of a certain density of object and linguistic phenomena across the corpus.

Four versions of each document are provided:

- Source text files (.sgm): All source files are encoded in UTF-8. These files use UNIX-style end of lines. Only text between the begin text tag <TEXT> and end text tag </TEXT> are to be evaluated. The one exception to this rule is that one TIMEX2 annotation is placed between the <DATETIME> and </DATETIME> tags, even though they occur outside the TEXT tags, in order to provide an anchor for time references within the text of the document.
- ACE Program Format files (.apf.xml).
- LDC Annotation Graph Format files (.ag.xml). AG is the LDC's internal annotation file format for ACE. These files can be viewed with the LDC's annotation tool[7].
- TABLE files (.tab): These files store mapping tables between the IDs used in each ag.xml file and their corresponding apf.xml file.

To verify data format integrity, three DTD's are distributed with the ACE local tasks training corpus. One DTD is used to verify the APF format, one to verify the AG format, and one to verify the original source document format. Appendix C contains the DTD used for the global tasks.

Table 6 ACE training corpus statistics for release LDC2007E63.

Source	Training epoch	Approximate size
English Resources		
Broadcast News	3/03 – 6/03	55,000 words
Broadcast Conversations	3/03 – 6/03	40,000 words
Newswire	3/03 – 6/03	50,000 words
Weblog	11/04 – 2/05	40,000 words
Usenet	11/04 – 2/05	40,000 words
Conversational Telephone Speech	11/04-12/04 (differentiated by topic vs. eval)	40,000 words
Arabic Resources		
Broadcast News	10/00 – 12/00	30,000+ words
Newswire	10/00 – 12/00	55,000+ words
Weblog	11/04 – 2/05	20,000+ words

3.2 The 2008 Evaluation Corpus

The evaluation corpus for ACE08 will be entirely new.

The evaluation source data for 2008 will include material from a variety of sources in English and Arabic. Selection of the source documents will be targeted to include a minimum number of occurrences of each type and subtype (for class "specific"), as well as certain linguistic phenomena. The latter represent various referential challenges for entity and relation mentions. For instance, interesting entity mentions would be orthographic and name variants, misspellings, nicknames, and aliases.

The characteristics of the ACE08 evaluation corpus have not been fully determined.

A key part of system output is the specification of entity mentions in terms of word locations in the source text. Word and phrase location information is specified in terms of the indices of the first and last characters of the word or phrase. ACE systems must compute these indices from the source data. Indices start with index 0 being assigned to the first character of a document. Ancillary information and annotation, which is provided as bracketed SGML tags, is not included in this count. Only characters (including white-spaces) outside of angle-bracketed expressions contribute to the character count. Also, each new line (nl or cr/lf) counts as one character.

4. Evaluation

Evaluation of ACE system performance will be supported for the entity and relation tasks defined in Section 2.

For each task and language combination chosen, **all source material must be processed** by the system being evaluated, including all of the different source types contained in the evaluation corpus.

Performance on each of the different ACE tasks is measured separately.

A total of 8 different evaluations will be available. These are listed in Table 7.

[6] Registered participants will be contacted by the LDC with instructions on how to obtain the ACE 2008 training corpus.

[7] The LDC Annotation Graph Toolkit is available for download at http://projects.ldc.upenn.edu/ace/tools/2005Toolkit.html.

Table 7 Evaluations for ACE08

Task	AR	EN
Within-Document Co-reference Tasks		
Entity Detection & Recognition (EDR)	X	X
Relation Detection & Recognition (RDR)	X	X
Cross-Document Co-reference Tasks		
Within-Language Global EDR	X	X
Within-Language Global RDR	X	X

4.1 EVALUATION METHOD

System performance on each of the tasks is scored using a model of the application value of system output. This overall value is the sum of the value for each system output object, accumulated over all system outputs. The value of a system output is computed by comparing its attributes and associated information with the attributes and associated information of the reference that corresponds to it. When system output information differs from that of the reference, value is lost. And when system output is spurious (i.e., there is no corresponding reference), negative value typically results. Perfect system output performance is achieved when the system output matches the reference without error. The overall score of a system is computed as the system output information relative to this perfect output.[8] Detail of the valuation of system output and scoring is given in Scoring Formulas. Note that for GEDR scoring where mentions are not required, the mutual mention value (MMV) is set equal to 1.

The correspondence between reference and system output objects is determined automatically by a mapping algorithm that chooses the best one-to-one mapping of reference to system objects. The definition of "best" mapping previously has been the mapping that gives the highest score. However, with the change in the definition of the EDR value score (from mention-weighted to level-weighted scoring) it sometimes (rarely) occurs that a mapping that maximizes the score can be extremely counterintuitive. For this reason, the mapping that will be used for ACE08 to determine correspondence between reference and system output entities will always be the mention-weighted score, regardless of how the official scoring is performed.

4.2 EVALUATION TASKS

4.3 LOCAL EDR (AND EMD)

The EDR task is to detect (infer) ACE-defined entities from mentions of them in the source language and to recognize and output selected entity attributes and information about these entities, including information about their mentions. Among other things, this requires that all of the mentions of an entity be correctly associated with that entity. The Value of a system output entity is defined as the product of two factors that represent how accurately the entity's attributes are recognized and how accurately the entity's mentions are detected:

$$Value_{sys\_entity} = Entity\_Value(sys\_entity) \cdot Mentions\_Value(\{sys\_mentions\})$$

Refer to appendix A for a complete description of the EDR *Value* formula.

The EMD value formula is identical to that for EDR. For EMD, however, each entity mention is promoted to "entity" status, separately from other mentions, and thus becomes an entity with only one mention.

4.4 LOCAL RDR (AND RMD)

The RDR task is to detect (infer) ACE-defined relations within the source language and to recognize and output selected attributes and information about these relations, including information about their mentions and arguments. A major part of correctly detecting relations is correctly recognizing the arguments that are related by the relation. Therefore, good argument recognition performance is important to achieving good RDR performance. The value of a system output relation is defined as the product of two factors that represent how accurately the relation's attributes are recognized and how accurately the relation's arguments are detected and recognized:

$$Value_{sys\_relation} = Relation\_Value(sys\_relation) \cdot Arguments\_Value(\{sys\_arguments\})$$

Refer to appendix A for a complete description of the RDR *Value* formula.

RMD is a derivative task that supports evaluation of relation mentions. In RMD, each relation mention, for both system output and reference relations, is promoted to "relation" status and becomes a separate and independent relation and is then evaluated as in RDR. There are several differences between mapping and scoring for RMD and RDR, however. This stems from an inherent ambiguity in specifying the mentions of relation arguments, because often times there are several possible choices. This ambiguity is handled in the following way:

- System output argument mentions are promoted to separate independent argument elements (including entities and times). Reference argument mentions are not promoted and are left unchanged as mentions of larger elements. This allows a system argument mention to map to any of the reference argument mentions.

Two other differences between RMD and RDR scoring provide the desired RMD score characteristics:

- Positive overlap is required between reference and system output "extents", defined as the span of their Arg-1/Arg-2 mention heads.
- Argument values are defined to be 1 if the arguments are mappable, 0 otherwise. (A system argument is "mappable" if it has a non-null score with the corresponding reference argument.)

4.5 GLOBAL EDR

Global EDR will be evaluated over the evaluation subset of documents using the same value formulas and similar parameters[9] as local EDR. Entity ID's must be globally reconciled, so that the same global ID is used to identify the same entity when that entity is mentioned in different documents. Mapping between reference

[8] Historically, it has been found that loss of value is attributable mostly to misses (where a reference has no corresponding system output) and false alarms (where a system output has no corresponding reference). To a lesser extent, value is lost due to errors in determining attributes and other associated information in those cases where the system output has a corresponding reference object.

[9] As mentioned elsewhere, for GEDR scoring the "mutual mention value" and "Mentions_Value" parameters will be set equal to 1.

and system output entities will be global, and value for each entity will accrue over all documents in which the reference entity and/or its corresponding system output entity is mentioned. Note that for GEDR scoring where mentions are not required, Mentions_Value (see appendix A) is assigned a value of 1.

4.6 GLOBAL RDR

Global RDR will be evaluated over the evaluation subset of documents by comparing a system output list of documents for each unique ACE relation to a reference list of documents for that relation. A unique relation is defined by the relation type, subtype, and a pair of entity arguments. The REFID of a global entity must be used as the relation argument for at least one of the relation arguments. Mapping between reference and system output entities will be global, and value for each relation will accrue over all documents in which the reference relation and/or its corresponding system output relation is mentioned.

4.7 2008 EVALUATION AND SCORING CONDITIONS

ACE08 will use two separate scoring mechanisms, in order to diversify the assessment of system output. In addition to the value model that has been used in previous evaluations, the B-cubed algorithm[10] for scoring mention co-referencing will also be used. The ACE value model, as defined and used for the 2007 evaluation, will be considered primary. The B-cubed score will provide a supplemental means to explore co-reference resolution performance

4.7.1 VALUE SCORING

Each document contributes separately and independently to the value score. This means that each ACE target (entity or relation) will contribute to the score for each document that mentions that target, as defined in the appendix A. For example, if an entity is mentioned in N different documents, then that entity will have N separate value contributions, one for each of the N documents.

4.7.2 B-CUBED SCORING

The B-cubed scoring algorithm computes mention co-reference over all entity mentions, irrespective of document boundaries. Thus, for example, if an entity is mentioned in two different documents according to the reference key, then the system output for that entity must also include the mentions for both documents in order to achieve perfect precision and recall.

4.8 RULES

- **Use of the ACE08 evaluation test set (source or reference) for any purpose other than the official ACE evaluation is prohibited.**
- **Human examination of the test data before system hypotheses are submitted for evaluation is prohibited**.
- **No changes are allowed to the system once the evaluation data has been released.** Adaptive systems may, of course, change themselves in response to the source data that they are processing.
- **No human intervention is allowed during processing of the evaluation data**, or prior to the submission of your test site's results to NIST.[11] This means that, in addition to **disallowing modifications to your system**, there also must be **no modifications to the test data, or human examination of it**.
- For each evaluation combination of task, language, and processing mode for which system output is submitted, **all of the documents from all of the sources must be processed for that evaluation combination.**
- Sites will receive the evaluation source data from NIST (see section 5.3 Schedule) and must return results to NIST within the specified period.
- Every participating site must submit a detailed system description to NIST by June 30th, 2008, as defined in section 5.6.
- Every participating site must attend the evaluation workshop and present a system talk.

5. TOOLS AND PROCEDURES

5.1 XML VALIDATION TOOLS

A java implementation of an XML validator[12] is available from the NIST ACE web site. The XML validator will verify that a system output file conforms to the current ACE DTD.[13]

Before sites submit their system results to NIST for scoring, they **must** validate the results file using the XML validation tool and the current ACE APF DTD. ***Results that are not validated will not be accepted.***

5.2 ACE EVALUATION SOFTWARE

The ACE evaluation software is available for download from the NIST ACE web site.[14] This tool can be used as a development aid for all the ACE tasks defined for ACE05 and ACE07 (entities, relations, times, values, and events). Although evaluation in 2008 will concentrate on entities, times, and relations, developers can work on all aspects of ACE using the current scoring software. The scoring formulas are documented in appendix A.

5.3 SCHEDULE

Evaluation will proceed in two stages. The initial stage will involve scoring system output against ground truth annotations prepared prior to the release of the evaluation data. The results of this scoring stage for each individual site will be released to that site within a couple weeks of submission to NIST. The second stage will involve closer scrutiny of additional results, based on pooling data submitted by multiple systems. This stage will take several months, and the results will be released prior to the ACE08 evaluation workshop.

Table 8 gives the evaluation schedule.

[10] http://www.nist.gov/speech/tests/ace/2008/doc/scoring-paper.ps, Bagga, Amit and Breck Baldwin. 1998. "Algorithms for scoring coreference chains", Proceedings of the First International Conference on Language Resources and Evaluation workshop on Linguistic Coreference.

[11] It sometimes happens that a system bug is discovered during the course of processing the test data. In such a case, please consult with NIST via email (ace_poc@nist.gov) for advice. NIST will advise you on how to proceed. Repairs may be possible that allow a more accurate assessment of the underlying performance of a system. If this happens, modified results may be accepted, provided that a written explanation of the modification is submitted and provided that the original results are also submitted and documented.

[12] http://www.nist.gov/speech/tests/ace/2008/software.html

[13] The DTDs used for the ACE program, can be found at: http://www.nist.gov/speech/tests/ace/2008/doc.

[14] The ACE evaluation tools may be accessed from http://www.nist.gov/speech/tests/ace/2008/software.html.

Table 8 ACE 2008 Evaluation Schedule

Date	Event
March 15, 2008	Final version of the cross-document pilot corpus available for exploration of the XDOC tasks
April 25, 2008	Evaluation registration deadline
May 9-23, 2008	Evaluation period
June 30, 2008	System Description due
Jun/Jul 2008	Post-hoc adjudication of system output for XDOC tasks
Sept. 4-5, 2008 (tentative)	*Evaluation Workshop to be located in the Baltimore/Washington area (tentative)*

5.4 Submission of System Output to NIST

All ACE-2008 system submissions must be packaged by the follow specifications.

5.5 Packaging Your System Output

Note, that in many cases a system output file will contain results for more than one task (i.e. EDR and RDR). In such a case the exact same set of files should be copied to the EDR and RDR subdirectories as defined below.

STEP1: Create a top level directory for each of the ***languages*** attempted (Arabic | English):

Example: $> mkdir arabic english

STEP2: Create a subdirectory identifying the ***tasks*** attempted (LEDR | LRDR| GEDR | GRDR):

Example: $> mkdir english/ledr english/lrdr arabic/ledr

STEP3: In each of these subdirectories make one directory for each system submitted (choose a name that identifies your site, BBN, SHEF, SRI…):

Example: $> mkdir english/ledr/NIST1_primary

STEP4: Deposit all system output files in the appropriate system directory.

STEP5: Create a compressed tar file of your results and transfer them to NIST by FTP (ftp://jaguar.ncsl.nist.gov/incoming). After successful transmission send e-mail to ace_poc@nist.gov identifying the name of the file submitted. Alternatively you may send the compressed tar file directly to ace_poc@nist.gov .

5.6 System Description

A valuable tool in discovering the strengths and weakness of different algorithmic approaches is to use system descriptions.

Each participant must prepare a ***detailed*** system description covering each system submitted. System descriptions are due at NIST no later than 06/30/08.

System descriptions will be distributed to each participant before the evaluation workshop.

Each system description should include:

- The ACE tasks and languages processed
- Identification of the primary system for each task
- A description of the system (algorithms, data, configuration) used to produce the system output
- How contrastive systems differ from the primary system
- A description of the resources required to process the test set, including CPU time and memory
- Applicable references

A system description template is available.[15]

6. Guidelines for Publications

NIST Speech Group's HLT evaluations have been moving towards an open model which promotes interchange with the outside world. The rules governing the publication of ACE08 evaluation results are given in section 4.2.

6.1 NIST Publication of Results

At the conclusion of the evaluation cycle, NIST plans to produce a hard-bound proceeding which documents the evaluation. Participants will be given the opportunity to contribute ACE 2008 evaluation papers. In addition, a report will be posted on the NIST web space and will identify the participants and official ACE value scores achieved for each combination of task and language. Scores will be reported for the overall test set and for the different data sources.

The report that NIST creates should not be construed or represented as endorsements for any participant's system or commercial product, or as official findings on the part of NIST or the U.S. Government.

6.2 Participant's Publication of Results

Participants must refrain from publishing results and/or releasing statements of performance until the official ACE08 results are posted by NIST on approximately Sept. 30th, 2008.

Participants may not compare its results with the results of other participants, such as stating rank ordering or score difference. Participants will be free to publish results for their own system, but, **sites will not be allowed to name other participants, or cite another site's results without permission from the other site**. Publications should point to the NIST report as a reference[16].

All publications must contain the following NIST disclaimer:

> ***NIST serves to coordinate the ACE evaluations in order to support Automatic Content Extraction research and to help advance the sate-of-the-art in content extraction technologies. ACE evaluations are not viewed as a competition, as such results reported by NIST are not to be construed, or represented, as endorsements of any participant's system, or as official findings on the part of NIST or the U.S. Government***

Linguistic resources used in building ACE systems should be referenced in the system description. Corpora should be given a formal citation, like any other information source. LDC corpus references should adopt the following citation format:

> Author(s), Year. Catalog Title (Catalog Number). Linguistic Data Consortium, Philadelphia.

[15] http://www.nist.gov/speech/tests/ace/2008/doc/template_sys_desc.txt

[16] This restriction exists to ensure that readers concerned with a particular system's performance will see the entire set of participants and tasks attempted by all researchers.

For example:

> Christopher Walker, et al., 2006. ACE 2005 Multilingual Training Corpus (LDC2006T06). Linguistic Data Consortium, Philadelphia.

APPENDIX A – SCORING

EDR scoring

The EDR value score for a system is defined to be the sum of the values of all of the system's output entity tokens, normalized by the sum of the values of all reference entity tokens. The maximum possible EDR value score is 100 percent.

$$EDR\_Value_{sys} \quad = \quad \sum_i value\_of\_sys\_token_i \Bigg/ \sum_j value\_of\_ref\_token_j$$

The value of each system token is based on its attributes and on how well it matches its corresponding reference token. A globally optimum correspondence between system and reference tokens which maximizes *EDR_Value* is determined and used, subject to the constraint of one-to-one mapping between system and reference tokens.[17] The value of a system token is defined to be the difference of two value terms, one that is in accord with the reference token and one that is not. In this formula, *Mentions_Value*(*sys*,*ref*) and *Mentions_Value*($\overline{sys}$,*ref*) respectively measure the value of those system mentions that do and that don't correspond to reference mentions.

$$\begin{aligned} Value(sys, ref) \quad &= \quad Element\_Value(sys, ref) \cdot Mentions\_Value(sys, ref) \\ &\quad - W_{FA} \cdot Element\_Value(sys) \cdot Mentions\_Value(\overline{sys}, ref) \end{aligned}$$

Element_Value is a function of the attributes of the system token and, if mapped, how well they match those of the corresponding reference token. In particular, *Element_Value* is defined as the product of the values of the token's attributes, specifically the token's ***type***, ***subtype***, ***class***, and ***names***. This value is then reduced for any attribute errors for the attributes ***type***, ***subtype***, ***class***, and ***names***, using the attribute error weighting parameters, $\{W_{err\text{-}attribute}\}$.

$$Element\_Value(sys, ref) \quad = \quad \prod_{\substack{attribute= \\ type,subtype,class,names}} \left\{ \min\begin{pmatrix} AttrValue(attribute_{sys}) \\ AttrValue(attribute_{ref}) \end{pmatrix} \cdot W_{err-attribute} \right\}$$

$$Element\_Value(sys) \quad = \quad \prod_{\substack{attribute= \\ type,subtype,class,names}} AttrValue(attribute_{sys})$$

Because names require a more complex comparison than the other attributes, $W_{err\text{-}names}$ is a function rather than a mere constant.

$$W_{err-names} = \frac{\displaystyle\sum_{all\ unique\ ref\ names} \max_{all\ unique\ sys\ names} \left(similarity(name_{ref}, name_{sys})\right) \; + \sum_{all\ unique\ sys\ names} \max_{all\ unique\ ref\ names} \left(similarity(name_{ref}, name_{sys})\right)}{\displaystyle\sum_{all\ unique\ ref\ names}(1) \; + \sum_{all\ unique\ sys\ names}(1)}$$

$$\text{where } similarity(string1, string2) \;=\; 1 \;-\; \text{levenshtein\_distance}(string1, string2) / \max(length(string1), length(string2))$$

Mentions_Value is a function of the mutual mention value (*MMV*) between the mentions of the system token and, if mapped, those of the corresponding reference token. A mention's *MMV* depends on the mention's ***type*** value parameter, *MTypeValue*, with this value being reduced for any errors in the mention attributes ***type***, ***role,*** and ***style***, using the mention attribute error weighting parameters, $\{W_{Merr}\}$.

$$MMV(mention_{sys}, mention_{ref}) \quad = \quad \begin{cases} \min\begin{pmatrix} MTypeValue(mention_{sys}), \\ MTypeValue(mention_{ref}) \end{pmatrix} \cdot \displaystyle\prod_{\substack{attribute= \\ type,role,style}} W_{Merr-attribute} & \text{if } mention_{sys} \text{ and } mention_{ref} \text{ correspond} \\ 0 & \text{otherwise} \end{cases}$$

$$MMV(mention_{sys}) \quad = \quad MTypeValue(mention_{sys})$$

For each pairing of a system token with a reference token, an optimum correspondence between the mentions of the system and reference tokens is determined. This mapping maximizes *Mentions_Value*, subject to the constraint of one-to-one mapping between system and reference mentions.

Mentions_Value is computed using one of two formulas, depending on whether valuation is ***mention***-weighted or ***level***-weighted. For mention-weighted valuation *Mentions_Value* is simply the sum of *MMV* over all mentions in all documents. For level-weighted valuation *Mentions_Value* is determined by a system token's ***level***[18] (and the level of its corresponding reference token), by the degree of correspondence between system and reference mentions, and by the number of documents in which the token is mentioned.

[17] System tokens and reference tokens are permitted to correspond only if they each have at least one mention in correspondence (for local EDR) or at least one document in common (for global EDR).

[18] A document entity's ***level*** is the highest (the most valued) ***type*** of mention that is used to refer to that entity in the document, and the ***level*** attribute value is equal to the mention ***type*** value for that level: *AttrValue*(*level*) = *MTypeValue*(*level*). (However, if the ***style*** of a mention is metonymic, then that mention's ***type*** is limited to NOM for determining the level of the entity in that document.)

For mention-weighted scoring, *Mentions_Value* is:

$$Mentions\_Value(sys, ref) = \sum_{\substack{all\\ docs}} \left(\sum_{\substack{all\ sys\ mentions\\ in\ doc\ that\ map\\ to\ ref\ mentions}} MMV\left(mention_{sys}, mention_{ref}\right) \right)$$

$$Mentions\_Value(\overline{sys}, ref) = \sum_{\substack{all\\ docs}} \left(\sum_{\substack{all\ sys\ mentions\\ in\ doc\ that\ \mathbf{don't}\\ map\ to\ ref\ mentions}} MMV\left(mention_{sys}\right) \right)$$

For level-weighted scoring, *Mentions_Value* is:

$$Mentions\_Value(sys, ref) = MTypeValue\left(level_{doc,ref}\right) \cdot \sum_{\substack{all\\ docs}} \left(\sum_{\substack{all\ sys\ mentions\\ in\ doc\ that\ map\\ to\ ref\ mentions}} MMV\left(mention_{sys}, mention_{ref}\right) \Big/ \sum_{\substack{all\ ref\ mentions\\ in\ doc}} MMV\left(mention_{ref}\right) \right)$$

$$Mentions\_Value(\overline{sys}, ref) = MTypeValue\left(level_{doc,sys}\right) \cdot \sum_{\substack{all\\ docs}} \left(\sum_{\substack{all\ sys\ mentions\\ in\ doc\ that\ \mathbf{don't}\\ map\ to\ ref\ mentions}} MMV\left(mention_{sys}\right) \Big/ \sum_{\substack{all\ sys\ mentions\\ in\ doc}} MMV\left(mention_{sys}\right) \right)$$

System mentions and reference mentions are permitted to correspond only if their **heads** have a mutual overlap of at least *min_overlap* and the text of their **heads** share a (fractional) consecutive string of characters[19] of at least *min_text_match*. Mention regions and overlaps are measured in terms of *number of characters* for text input, in terms of *time* for audio input, and in terms of *area* for image input.

$$mutual\_overlap = \frac{sys\_head \bigcap ref\_head}{\max\left(sys\_head, ref\_head\right)}$$

$$fractional\_consecutive\_string = \frac{\left(\begin{array}{c} \#\ of\ characters\ in\ the\ longest\ consecutive\ string\ of\ characters \\ that\ is\ contained\ in\ both\ system\ and\ reference\ mention\ head\ texts \end{array}\right)}{\max\left(\begin{array}{l} \#\ of\ characters\ in\ system\ mention\ head\ text, \\ \#\ of\ characters\ in\ reference\ mention\ head\ text \end{array}\right)}$$

The current default scoring parameters for EDR are given in Table 4.

Table 4 Default parameters for scoring EDR performance

W_{FA} = 0.75							
Element_Value **parameters**				***Mentions_Value*** **parameters**			
Attribute	$W_{err\text{-}attribute}$	**Attribute Value**	***AttrValue***	**Attribute**	$W_{Merr\text{-}attribute}$	**Attribute Value**	***MTypeValue***
Type	0.50	(all types)	1.00	Type	0.90	NAM	1.00
Class	0.75	SPC	1.00			NOM	0.50
		(not SPC)	0.00			PRO	0.10
Subtype	0.90	(all types)	1.00	Role	0.90	n/a	n/a
Name	use formula[20]	(all types)	1.00	Style	0.90	n/a	n/a
Mapping = mention-weighted		***Valuation = level-weighted***		***min_overlap*** = 0.30		***min_text_match*** = 0.00	

[19] This requirement of a common substring in both system and output mention heads was invoked to account for errors in transcribing speech and image data into text. The intent is to require a mention to be meaningful and relevant in order to be counted.

[20] For official 2008 scoring, only GEDR will use the name similarity formula. For LEDR scoring, $W_{err\text{-}names} = 1$.

RDR scoring

The RDR value score for a system is defined to be the sum of the values of all of the system's output relation tokens, normalized by the sum of the values of all reference relation tokens. The maximum possible RDR value score is 100 percent.

$$RDR\_Value_{sys} \;=\; \sum_i value\_of\_sys\_token_i \Big/ \sum_j value\_of\_ref\_token_j$$

The value of each system token is based on its attributes and arguments and on how well they match those of a corresponding reference token. A globally optimum correspondence between system and reference tokens which maximizes *RDR_Value* is determined and used, subject to the constraint of one-to-one mapping between system and reference tokens. System tokens and reference tokens are permitted to correspond only if they have some nominal basis for correspondence. The required nominal basis is selectable from the set of minimal conditions listed in Table 5.

Table 5 Conditions required for correspondence between system and reference relation tokens

Condition	Description
arguments	At least one argument in the system token must be mappable to an argument in the reference token.
extents	The system and reference tokens must each have at least one mention extent in correspondence with the other.
both	Both the **arguments** condition and the **extents** condition must be met.
either	Either the **arguments** condition or the **extents** condition must be met.
all	All arguments in the reference token must be one-to-one mappable to arguments in the system token.
all+extents	Both the **all** condition and the **extents** condition must be met.

The value of a system token is defined by the following formula:

$$\begin{aligned} Value(sys,ref) \;=\; & Element\_Value(sys,ref) \cdot Arguments\_Value(sys,ref) \\ & -W_{FA} \cdot Element\_Value(sys) \cdot Arguments\_Value(\overline{sys},ref) \end{aligned}$$

In this expression for the *Value* of a system token, *Arguments_Value*(*sys*,*ref*) and *Arguments_Value*($\overline{sys}$,*ref*) respectively measure the value of system arguments that do and that don't correspond to reference arguments.

Element_Value is a function of the attributes of the system token and, if mapped, how well they match those of the corresponding reference token. The inherent value of a token is defined as the product of the token's attribute value parameters, {*AttrValue*}, for the attributes ***type*** and ***modality***. This inherent value is reduced for any attribute errors (i.e., for any difference between the values of system and reference attributes), using the error weighting parameters, $\{W_{err\text{-}attribute}\}$. If a system token is unmapped, then the value of that token is weighted by a false alarm penalty, W_{FA}.

$$Element\_Value(sys,ref) \;=\; \prod_{\substack{attribute= \\ type,modality}} \min\begin{pmatrix} AttrValue(attribute_{sys}) \\ AttrValue(attribute_{ref}) \end{pmatrix} \cdot \prod_{\substack{attribute= \\ type,subtype,modality,tense}} W_{err-attribute}$$

$$Element\_Value(sys) \;=\; \prod_{\substack{attribute= \\ type,modality}} AttrValue(attribute_{sys})$$

Arguments_Value is a function of the mutual argument value (*MAV*) between the arguments of the system token and, if mapped, those of the corresponding reference token. An argument's *MAV*, if mapped, is equal to the mapped value of the elements serving as arguments, *Value*(arg_{sys},arg_{ref}), but reduced in value if the system's argument role is in error.

$$MAV(arg_{sys},arg_{ref}) \;=\; Value(arg_{sys},arg_{ref}) \cdot W_{err-role} \cdot W_{err-asym} \quad \text{if } Mentions\_Value(arg_{sys},arg_{ref}) \;>\; 0$$

$$MAV(arg_{sys}) \;=\; Value(arg_{sys},arg_{sys})$$

There are several requirements that must be satisfied in order for a reference argument to be considered to be in correspondence to a system argument. First, note that there are two required arguments, namely the two arguments for which the relation is being asserted. These arguments have roles called "Arg-1" and "Arg-2", and there may be only one Arg-1 and one Arg-2 argument.[21] The requirements for correspondence are listed in Table 6.

Table 6 Conditions required for correspondence between system and reference relation arguments

Condition	**Requirement**
Always	The reference argument must be mappable to the system argument. That is, they must have at least one mention in correspondence.
Argument role is Arg-1 or Arg-2 and the relation symmetric	The reference argument role may be either "Arg-1" or "Arg-2", and no role mismatch penalty is imposed.
Argument role is Arg-1 or Arg-2 and the relation is **not** symmetric	The reference argument role may be either "Arg-1" or "Arg-2", but an asymmetry error penalty, $W_{err\text{-}asym}$, is imposed.
If the "**mapped**" argument option is invoked	The reference argument must correspond to the system argument. That is, they must be mapped to each other at the argument level.

For each pairing of a system relation token with a reference relation token, an optimum correspondence between system arguments and reference arguments that maximizes *Arguments_Value* is determined and used. This optimum mapping is constrained to be a one-to-one mapping between system and reference arguments.

Arguments_Value is computed using the following formula:

$$Arguments\_Value(sys, ref) = \sum_{\substack{\text{all } arg_{sys} \text{ with a} \\ \text{corresponding } arg_{ref}}} \left(\sum_{\substack{\text{all docs that} \\ \text{mention the relation}}} MAV_{doc}\left(arg_{sys}, arg_{ref}\right) \right)$$

$$Arguments\_Value(\overline{sys}, ref) = \sum_{\substack{\text{all } arg_{sys} \text{ with } \textbf{no} \\ \text{corresponding } arg_{ref}}} \left(\sum_{\substack{\text{all docs that} \\ \text{mention the relation}}} MAV_{doc}\left(arg_{sys}, arg_{sys}\right) \right)$$

The current default scoring parameters for RDR are given in Table 7.

Table 7 Default parameters for scoring RDR performance

W_{FA} = 0.75			
Element_Value parameters			**Arguments_Value parameters**
Relation mapping requirements (Table 5) = "arguments"			**"mapped"** *arguments optional requirement* **NOT** *invoked* (Table 5)
attribute	**AttrValue**	**$W_{err\text{-}attribute}$**	
type	**1.00 for all types**	**1.00**	**Both** Arg-1 and Arg-2 arguments must be mappable (i.e., must have non-null *MAV*'s)
modality	**1.00 for all types**	**0.75**	
subtype	**(not applicable)**	**0.70**	$W_{err\text{-}role}$ = 0.75
tense	**(not applicable)**	**1.00**	**$W_{err\text{-}asym}$ = 0.70**

[21] Arg-1 and Arg-2 are the only roles for which the number of arguments is limited.

B-cubed scoring of entity mentions

The B-cubed scoring of entity mentions follows the algorithm described in Bagga and Baldwin's 1998 paper entitled "Algorithms for Scoring Coreference Chains". This algorithm produces a measure of how well system output mentions are clustered into entities without determining an explicit mapping of system output entities to reference entities. Specifically, for each pair of corresponding reference and system output mentions, the B-cubed algorithm computes the mention precision and recall of the host entities:

$$Precision(sys\_mention_i) = \frac{\left\{\begin{array}{l}\text{the maximum number of mentions in } sys\_entity_i \text{ that have corresponding ref entity} \\ \text{mentions, for ref entities that contain a mention that corresponds to } sys\_mention_i\end{array}\right\}}{\text{the number of mentions in } sys\_entity_i}$$

where sys_entity_i is the entity that contains $sys\_mention_i$, and

$$Recall(ref\_mention_j) = \frac{\left\{\begin{array}{l}\text{the maximum number of mentions in } ref\_entity_j \text{ that have corresponding sys entity} \\ \text{mentions, for sys entities that contain a mention that corresponds to } ref\_mention_j\end{array}\right\}}{\text{the number of mentions in } ref\_entity_j}$$

where ref_entity_j is the entity that contains $ref\_mention_j$.

These mention-specific precision and recall values are then averaged over all mentions to produce the B-cubed precision and recall:

$$Precision_{B-cubed} = \underset{all\ i}{average}\{Precision(sys\_mention_i)\}$$

$$Recall_{B-cubed} = \underset{all\ j}{average}\{Recall(ref\_mention_j)\}$$

In addition to computing the B-cubed precision and recall using a simple count of corresponding mentions, a value-weighted version of B-cubed precision and recall is also computed using the mutual mention value (*MMV*) between reference and system output mentions. The *MMV*, as described for EDR scoring, is a function of the mention type and is discounted for differences between the type, role and style of reference and system mentions. The *MMV* value-weighted versions of B-cubed precision and recall are:

$$Value\_Precision(sys\_mention_i) = \frac{\left\{\begin{array}{l}\text{the maximum of the sum of } MMV\text{'s between } sys\_entity_i \text{ and those} \\ \text{ref entities that contain a mention that corresponds to } sys\_mention_i\end{array}\right\}}{\text{the sum of the } MMV\text{'s between } sys\_entity_i \text{ mentions and themselves}}$$

$$Value\_Recall(ref\_mention_j) = \frac{\left\{\begin{array}{l}\text{the maximum of the sum of } MMV\text{'s between } ref\_entity_j \text{ and those} \\ \text{sys entities that contain a mention that corresponds to } ref\_mention_j\end{array}\right\}}{\text{the sum of the } MMV\text{'s between } ref\_entity_j \text{ mentions and themselves}}$$

These value-weighted mention-specific precision and recall values are then weighted by each mention's type value, as described for EDR scoring, and then averaged over all mentions to produce the value-weighted B-cubed precision and recall:

$$Value\_Precision_{B-cubed} = \sum_i MTypeValue(sys\_mention_i) \cdot Value\_Precision(sys\_mention_i) \Big/ \sum_i MTypeValue(sys\_mention_i)$$

$$Value\_Recall_{B-cubed} = \sum_j MTypeValue(ref\_mention_j) \cdot Value\_Recall(ref\_mention_j) \Big/ \sum_j MTypeValue(ref\_mention_j)$$

APPENDIX B – AN ANNOTATED SAMPLE OF SCORING OUTPUT FROM THE ACE EVALUATION TOOL

The primary evaluation output from the ACE evaluation tool comprises detection and recognition statistics that include basic count statistics and value statistics derived from the value model. These statistics are broken out according to various target and source attributes. Shown below are two examples of detection and recognition statistics, the first a breakout for entity statistics conditioned on entity TYPE, and the second a breakout for relation statistics conditioned on relation TYPE. There are 31 columns, labeled “AA” through “ZZ”, with a brief description of each of the columns following the examples.

Entity Detection and Recognition statistics:

```
Entity  _Entity_Count__    ___Document_Count____    _____Document_Count_(%)______     _________________________Cost_(%)_________________________     ____________Unconditioned_Cost_(%)____________
TYPE     Ref   Detection     Ref   Detection   Rec    Detection   Rec   B3 Unweighted     Detection  Attr   ___Mentions___    Value  B3 Value_based      Max     Detection    Attr     ____Mentions____
         Tot    FA  Miss     Tot    FA  Miss   Err     FA  Miss   Err    Pre--Rec--F       FA  Miss   Err    FA  Miss   Err     (%)    Pre--Rec--F      Value    FA    Miss    Err      FA   Miss    Err
   FAC   225    57    48     234    58    50    89    24.8  21.4  38.0   84.8 51.2 63.9    21.1  15.6  13.2   3.8  10.7   2.2    33.3   82.6 59.8 69.4     5.66   1.19   0.88   0.75    0.22   0.60   0.13
   GPE   275    32    20     577    69    43    82    12.0   7.5  14.2   86.7 65.4 74.6     7.7   3.9   3.0   2.8   6.4   2.2    74.0   86.3 78.9 82.4    27.72   2.13   1.08   0.83    0.77   1.77   0.62
   LOC   127    26    56     144    30    64    47    20.8  44.4  32.6   68.2 47.6 56.1    20.9  31.8  12.9   6.0   4.8   1.0    22.5   64.5 58.1 61.1     3.24   0.68   1.03   0.42    0.19   0.16   0.03
   ORG   514    94    92     619   113   111   193    18.3  17.9  31.2   74.3 58.6 65.5    14.2   9.5   4.6   4.8   6.9   1.0    59.0   79.5 74.2 76.8    22.52   3.20   2.13   1.04    1.07   1.56   0.22
   PER  1572   359   346    1791   408   395   621    22.8  22.1  34.7   79.7 60.9 69.0    24.5   5.6   1.7   6.6  12.9   0.6    47.9   87.3 74.6 80.5    36.78   9.02   2.08   0.64    2.44   4.73   0.23
   VEH   122    13    32     127    13    34    44    10.2  26.8  34.6   75.7 48.2 58.9    13.2  19.4   4.9   9.3  17.5   0.6    35.1   75.4 50.8 60.7     2.40   0.32   0.47   0.12    0.22   0.42   0.02
   WEA   135    15    57     135    15    57    46    11.1  42.2  34.1   81.7 43.6 56.9    17.9  36.5   2.2   5.7   7.5   0.4    29.9   79.8 48.5 60.3     1.67   0.30   0.61   0.04    0.09   0.12   0.01
 total  2970   596   651    3627   706   754  1122    19.5  20.8  30.9   79.8 59.9 68.5    16.8   8.3   3.8   5.0   9.4   1.3    55.4   84.8 73.6 78.8   100.00  16.85   8.27   3.82    5.01   9.37   1.26

    AA    AB    AC    AD       A     B     C     D       E     F     G      H    I    J       K     L     M     N     O     P       Q      R    S    T        U      V      W      X       Y      Z     ZZ
```

Relation Detection and Recognition statistics:

```
relation Relation_Count_   ___Document_Count____    _____Document_Count_(%)______     _________________________Cost_(%)_________________________     ____________Unconditioned_Cost_(%)____________
TYPE     Ref   Detection     Ref   Detection   Rec    Detection   Rec     Unweighted      Detection  Attr   __Arguments___    Value   Value-based       Max     Detection    Attr     ___Arguments____
         Tot    FA  Miss     Tot    FA  Miss   Err     FA  Miss   Err    Pre--Rec--F       FA  Miss   Err    FA  Miss   Err     (%)    Pre--Rec--F      Value    FA    Miss    Err      FA   Miss    Err
    ART  217    17   135     217    17   135    37     7.8  62.2  17.1   45.5 20.7 28.5     6.8  59.2   1.5   0.0   2.2  12.7    17.6   51.2 24.4 33.1     9.93   0.68   5.87   0.15    0.00   0.22   1.26
GEN-AFF  202    19    93     204    20    94    67     9.8  46.1  32.8   33.1 21.1 25.7     8.6  41.2   6.6   0.0   0.5  16.1    27.1   52.9 35.7 42.6    11.91   1.02   4.90   0.79    0.00   0.05   1.92
METONYM    9     0     9       9     0     9     0     0.0 100.0   0.0    0.0  0.0  0.0     0.0 100.0   0.0   0.0   0.0   0.0     0.0    0.0  0.0  0.0     0.58   0.00   0.58   0.00    0.00   0.00   0.00
ORG-AFF  401    28   181     438    30   198   128     6.8  45.2  29.2   41.5 25.6 31.6     5.3  43.0   4.3   0.0   1.3  15.8    30.4   57.3 35.7 44.0    25.96   1.37  11.15   1.12    0.00   0.32   4.09
PART-WH  298    15   190     315    15   201    37     4.8  63.8  11.7   59.7 24.4 34.7     3.7  61.4   1.0   0.0   0.0   9.4    24.5   66.7 28.2 39.7    18.92   0.69  11.62   0.19    0.00   0.00   1.78
PER-SOC  162    13    86     179    14    96    36     7.8  53.6  20.1   48.5 26.3 34.1     6.2  46.8   1.7   0.0   0.0  16.9    28.3   58.2 34.5 43.3     8.99   0.56   4.21   0.16    0.00   0.00   1.52
   PHYS  379    42   275     379    42   275    84    11.1  72.6  22.2   13.7  5.3  7.6     7.3  69.3   2.2   0.3   4.2  10.2     6.5   36.2 13.7 19.9    23.70   1.72  16.44   0.53    0.07   0.99   2.42
  total 1668   135   969    1741   138  1008   389     7.9  57.9  22.3   39.5 19.8 26.3     6.0  54.8   2.9   0.1   1.6  13.0    21.6   53.9 27.6 36.6   100.00   6.04  54.77   2.93    0.07   1.59  12.99

    AA    AB    AC    AD       A     B     C     D       E     F     G      H    I    J       K     L     M     N     O     P       Q      R    S    T        U      V      W      X       Y      Z     ZZ
```

AA) The condition for which performance in columns **A** through **ZZ** is computed

AB) The number of (global) reference elements

AC) The number of (global) system elements with no corresponding reference element (false alarms)

AD) The number of (global) reference elements with no corresponding system element (misses)

A) The number of document-level occurrences of reference elements

B) The number of document-level occurrences of system elements with no corresponding reference element (false alarms)

C) The number of document-level occurrences of reference elements with no corresponding system element (misses)

D) The number of document-level occurrences of reference elements that were detected by the system but recognized imperfectly

E) The number of document-level false alarms, expressed as a percentage of reference elements

F) The number document-level misses, expressed as a percentage of reference elements

G) The number of imperfectly recognized document-level element occurrences, expressed as a percentage of reference elements

H) *for entities:* B-cubed Precision
for relations: The percentage of document-level occurrences of system elements that were perfectly recognized

I) *for entities:* B-cubed Recall
for relations: The percentage of document-level reference elements that were perfectly recognized

J) The F-measure using (**H**) as precision and (**I**) as recall

K) The percentage value lost due to false alarms

L) The percentage value lost due to misses

M) The percentage value lost due to errors in recognizing element attributes

N) *for entities:* The percentage value lost due to spurious mentions
for relations: The percentage value lost due to spurious arguments

O) *for entities:* The percentage value lost due to missed mentions
for relations: The percentage value lost due to missed arguments

P) *for entities:* The percentage value lost due to mention recognition errors
for relations: The percentage value lost due to argument recognition errors

Q) **The “bottom line” value of the system output, after subtracting all of the costs (i.e., the lost value) due to detection and recognition errors.** This is the one overall score that is intended to represent the overall performance of the system.

R) *for entities:* B-cubed Value_Precision
for relations: The value of system output as a percentage of apparent system value

S) *for entities:* B-cubed Value_Recall
for relations: The value of system output as a percentage of reference value

T) The F-measure using (**R**) as precision and (**S**) as recall

U) The reference value of this condition as a percentage of total reference value

V) through **ZZ**) Unconditioned costs: These columns are the same as columns **K**) through **P**) except that they are expressed here as a percentage of the **total** reference value rather than as a percentage of the reference value for the given condition

APPENDIX C – A CONDENSED VERSION OF THE APF DTD (VER 5.2.0) TO BE USED FOR THE 2008 EVALUATION

```
<!ELEMENT source_file      (document+)>

<!ATTLIST source_file
                           URI      CDATA              #REQUIRED
                           SOURCE   CDATA              #IMPLIED
                           TYPE     (text|audio|image) #REQUIRED
                           VERSION  NMTOKEN            #IMPLIED
                           AUTHOR   CDATA              #IMPLIED
                           ENCODING CDATA              #IMPLIED
>

<!ELEMENT document         (entity|relation)* >
<!ATTLIST document
                           DOCID CDATA #REQUIRED
>

<!-- Entities -->
<!ELEMENT entity         (entity_mention*,entity_attributes*,external_link*)>
<!ATTLIST entity
                         ID      ID                               #REQUIRED
                         TYPE    (PER|ORG|LOC|GPE|FAC|VEH|WEA)      #REQUIRED
                         SUBTYPE (Individual|Group|Indeterminate|
                           Government|Non-Governmental|
                           Commercial|Educational|
                           Media|Religious|Sports|
                           Medical-Science|Entertainment|
                           Address|Boundary|Water-Body|Celestial|
                           Land-Region-Natural|Region-General|
                           Region-International|Continent|Nation|
                           State-or-Province|County-or-District|
                           Population-Center|GPE-Cluster|Special|
                           Building-Grounds|Subarea-Facility|Path|
                           Airport|Plant|Land|Air|Water|Subarea-Vehicle|
                           Blunt|Exploding|Sharp|Chemical|
                           Biological|Shooting|Projectile|Nuclear|
                           Underspecified)                        #REQUIRED
                         CLASS   (NEG|SPC|GEN|USP)                #REQUIRED
>

<!ELEMENT entity_attributes (name*)>

<!ELEMENT name             (bblist|charspan|charseq|timespan)?>
<!ATTLIST name             NAME  CDATA                           #REQUIRED
>

<!-- Entity Mentions
   Note:  entity mentions (entity_mention elements)
          are not required for scoring GEDR -->

<!ELEMENT entity_mention   (extent, head)>
<!ATTLIST entity_mention
                           ID        ID                         #REQUIRED
                           TYPE      (NAM|NOM|PRO)              #REQUIRED
                           LDCTYPE   (NAM|NOM|BAR|PRO|WHQ|
                                      HLS|PTV|APP|ARC|
                                      EAP|NAMPRE|NOMPRE|
                                      NOMPOST|NAMPOST)           #IMPLIED
                           ROLE      (PER|ORG|LOC|GPE)           #IMPLIED
                           METONYMY_MENTION (TRUE|FALSE)         #IMPLIED
                           LDCATR    (TRUE|FALSE)                #IMPLIED
>
```

```
<!-- Relations -->
<!ELEMENT relation           (relation_argument,
                              relation_argument+,
                              relation_mention*)>
<!ATTLIST relation
                             ID       ID                        #REQUIRED
                             TYPE     (PHYS|PART-WHOLE|PER-SOC|ORG-AFF|
                                       ART|GEN-AFF|METONYMY)    #REQUIRED
                             SUBTYPE  (Located|Near|Geographical|
                                       Subsidiary|Artifact|Business|
                                       Family|Lasting-Personal|Employment|
                                       Ownership|Founder|Student-Alum|
                                       Sports-Affiliation|
                                       Investor-Shareholder|
                                       Membership|
                                       User-Owner-Inventor-Manufacturer|
                                       Citizen-Resident-Religion-Ethnicity|
                                       Org-Location)            #IMPLIED
                             MODALITY (Asserted|Other)          #IMPLIED
                             TENSE    (Past|Present|Future|
                                       Unspecified)             #IMPLIED
>

<!ELEMENT relation_argument  EMPTY>
<!ATTLIST relation_argument
                             REFID     IDREF               #REQUIRED
                             ROLE      (Arg-1|Arg-2|
                                        Time-Within|
                                        Time-Starting|
                                        Time-Ending|
                                        Time-Before|
                                        Time-After|
                                        Time-Holds|
                                        Time-At-Beginning|
                                        Time-At-End)       #REQUIRED
>

<!-- Relation Mentions
   Note:  relation mentions (relation_mention elements) and relation
          mention arguments (relation_mention_argument elements)
          are not required for scoring either LRDR or GRDR -->

<!ELEMENT relation_mention  (extent,
                             relation_mention_argument,
                             relation_mention_argument+)>
<!ATTLIST relation_mention
                            ID                ID                #REQUIRED
                            LEXICALCONDITION  (Possessive|Preposition|
                                               PreMod|Formulaic|Verbal|
                                               Participial|Other|
                                               Coordination)    #IMPLIED
>

<!ELEMENT relation_mention_argument  (extent?)>
<!ATTLIST relation_mention_argument
                            REFID     IDREF                 #REQUIRED
                            ROLE      (Arg-1|Arg-2|
                                       Time-Within|
                                       Time-Starting|
                                       Time-Ending|
                                       Time-Before|
                                       Time-After|
                                       Time-Holds|
                                       Time-At-Beginning|
                                       Time-At-End)         #REQUIRED
>
```