ACE

Automatic Content Extraction

A program to develop technology to extract and characterize meaning from human language

Government ACE Team

- Project Management
  - NSA
  - CIA
  - DIA
  - NIST
- Research Oversight
  - JK Davis (NSA)
  - Charles Wayne (NSA)
  - Boyan Onyshkevych (NSA)
  - Steve Dennis (NSA)
  - George Doddington (NIST)
  - John Garofolo (NIST)

ACE Five-Year Goals

- Develop automatic content extraction technology to extract information from human language in textual form:

  Text (newswire) Speech (ASR) Image (OCR)

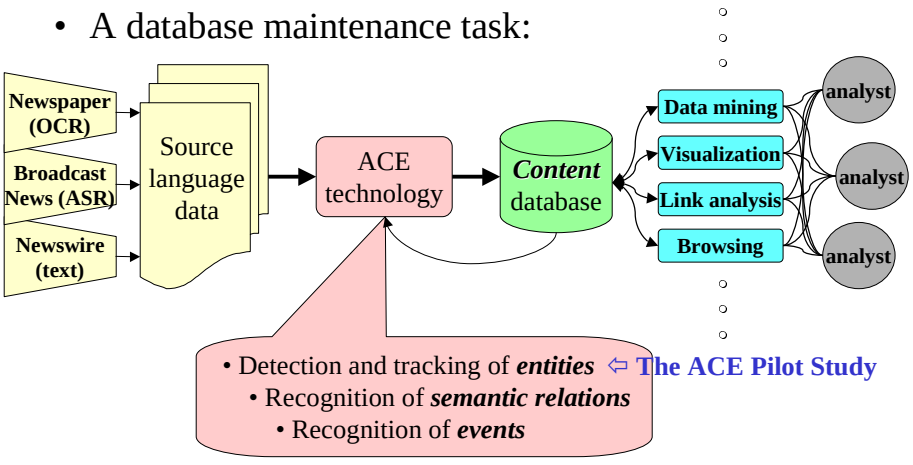
- Enable new applications in:

Data Mining	Browsing	Link Analysis
Summarization	Visualization	Collaboration
TDT	DR	IE

- Provide major improvements in analyst access to relevant data

The ACE Processing Model

- A database maintenance task:



The ACE Pilot Study

Objective: To lay the groundwork for the ACE program.

- Answer key questions:
  - What are the right technical goals?
  - What is the impact of degraded text?
  - How should performance be measured?
- Establish performance baselines
- Choose initial research directions (Entity Detection and Tracking)
- Begin developing content extraction technology

The ACE Pilot Study Process

- May ’99
  - Discuss/Explore candidate R&D tasks
  - Bimonthly meetings
  - Identify Data
  - Bimonthly site visits
  - Provide infrastructure support
    annotation / reconciliation / evaluation
  - Select/Define Pilot Study common task
  - Annotate Data
  - Implement and evaluate baseline systems
  - Final pilot study workshop (22-23 May ’00)
- May ’00

The Pilot Study R&D Task

***E**ntity **D**etection and **T**racking*
(limited to "within-document" processing)

EDT – a suite of four tasks:

1) Detection of Entities – limited to five types:
   PER ORG GPE LOC FAC
2) Recognition of Entity Attributes – limited to:
   Type
   Name
3) Detection of Entity Mentions
   (i.e., entity tracking)
4) Recognition of Mention Extent

The Entity Detection Task

- This is the most basic common task. It is the foundation upon which the other tasks are built, and it is therefore a required task for all ACE technology developers.
- Recognition of entity type and entity attributes is separate from entity detection. Note, however, that detection is limited to entities of specified types.

Entity Types

Entities to be detected and recognized will be limited to the following five types:

1. ***Person***. Person entities are limited to humans. A person may be a single individual or a group if the group has a group identity.
2. ***Organization***. Organization entities are limited to corporations, agencies, and other groups of people defined by an established organizational structure. Churches, schools, embassies and restaurants are examples of organization entities.
3. ***GPE*** *(A Geo-Political Entity)*. GPE entities are politically defined geographical regions. A GPE entity subsumes and does not distinguish between a geographical region, its government or its people. GPE entities include nations, states and cities.

Entity Types (continued)

4. ***Location***. Location entities are limited to geographic entities with physical extent. Location entities include geographical areas and landmasses, bodies of water, and geological formations. A politically defined geographic area is a GPE entity rather than a location entity.
5. ***Facility***. Facility entities are human-made artifacts falling under the domains of architecture and civil engineering. Facility entities include buildings such as houses, factories, stadiums, museums; and elements of transportation infrastructure such as streets, airports, bridges and tunnels.

The Entity Detection Process

- A system must output a representation of each entity mentioned in a document, at the end of that document:
  - Pointers to the beginning and end of the head of one or more mentions of the entity. (As an option, pointers to all mentions may be output, in order to support the evaluation of Mention Detection performance.)
  - Entity type and attribute (name) information.
  - Mention extent, in terms of pointers to the beginning and end of each mention. (optional – for evaluation of mention extent recognition performance only)

Evaluation of Entity Detection

Entity Detection performance will be measured in terms of missed entities and false alarm entities. In order to measure misses and false alarms, each reference entity must first be associated with the appropriate corresponding system output entity. This is done by choosing, for each reference entity, that system output entity with the best matching set of mentions. Note, however, that a system output entity is permitted to map to at most one reference entity.

- A ***miss*** occurs whenever a reference entity has no corresponding output entity.
- A ***false alarm*** occurs whenever an output entity has no corresponding reference entity.

Recognition of Entity Attributes

- This is the basic task of characterizing entities. It includes recognition of entity type. It is a required task for all ACE technology developers.
- *Performance is measured only for those entities that are mapped to reference entities.*
- Evaluation of performance will be conditioned on entity and attribute type.
- For the EDT pilot study, the only attributes to be recognized are entity type and entity name.
- An entity name is "recognized" by detecting its presence and then correctly determining its extent.

Detection of Entity Mentions

- Mention detection measures the ability of the system to correctly detect and associate all of the mentions of an entity, *for all correctly detected entities*. It is in essence a co-reference task.
- Detection performance will be measured in terms of missed mentions and false alarm mentions. For each mapped reference entity:
  - a ***miss*** occurs for each reference mention of that entity without a matching mention in the corresponding output entity, and
  - a ***false alarm*** occurs for each mention in the corresponding output entity without a matching reference mention.

Recognition of Mention Extent

- Extent recognition measures the ability of the system to correctly determine the extent of the mentions, *for all correctly detected mentions*.
- This ability will be measured in terms of the classification error rate, which is simply the fraction of all mapped reference mentions that have extents that are not "identical" to the extents of the corresponding system output mentions.

Action Items that remain to be completed for the ACE pilot study

- Annotate the Pilot Corpus
- ASR:
  - Publish ASR transcription output
  - Produce timing information for ref transcripts
- OCR:
  - Produce and publish OCR recognition output
  - Produce bounding boxes for ref transcripts
- EDT technology development:
  - Implement EDT systems
  - Evaluate them

The ACE/EDT Pilot Corpus

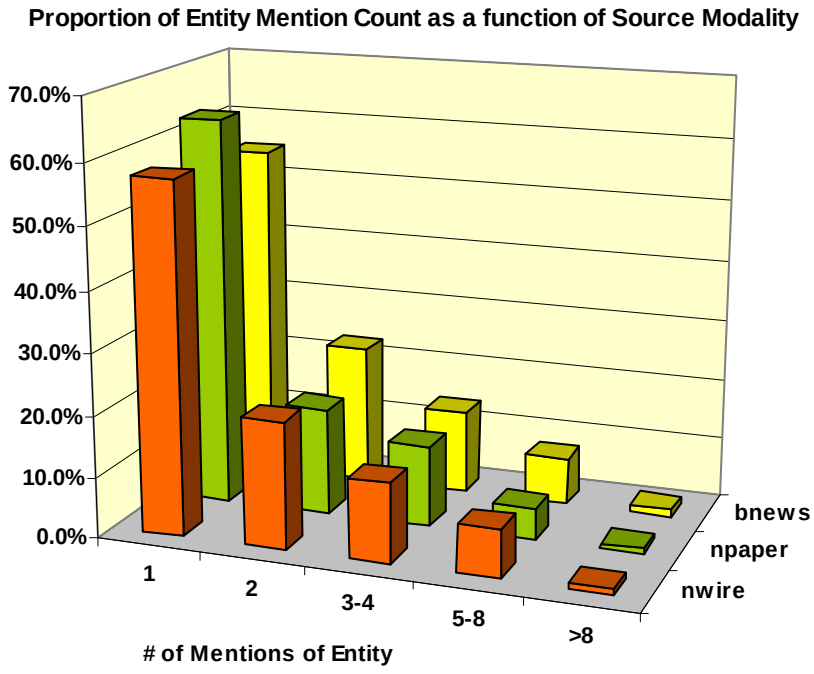
	Training 01-02/98	Dev Test 03-04/98	Eval Test 05-06/98
Newswire	30,000 words	15,000 words	15,000 words
Broadcast News	30,000 words	15,000 words	15,000 words
Newspaper	30,000 words	15,000 words	15,000 words

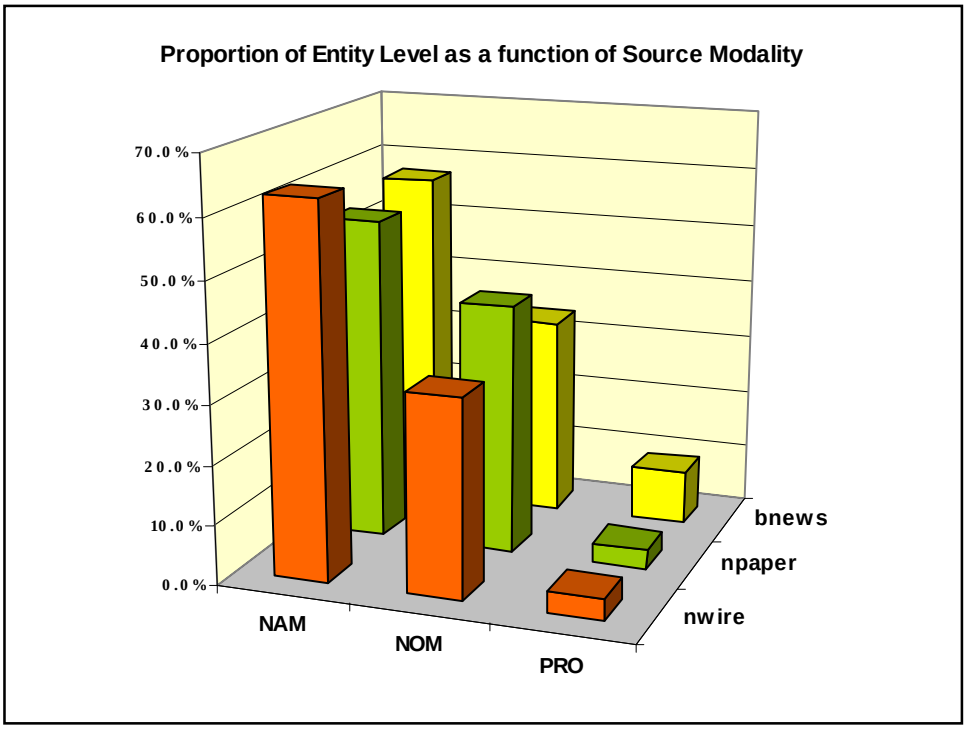
Schedule for Pilot Corpus Annotation and EDT Evaluation

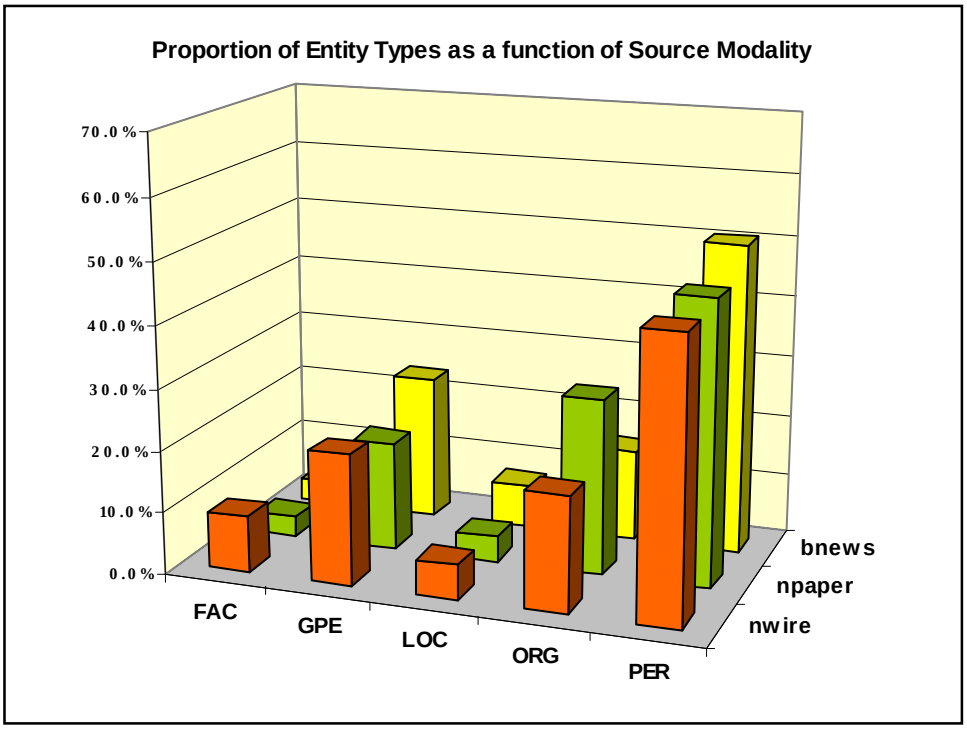
	Mon	Tue	Wed	Thu	Fri				
March	13	14	15	16	17	Training Data Annotation		Annotation sites make incremental releases of data	
	20	21	22	23	24				
	27	28	29	30	31		Nist releases trn data		
April	3	35	36	37	38	DevSet Data Annotation			
	10	11	12	13	14		Nist releases trn data		
	17	18	19	20	21	EvalSet Data Annotation			
	24	25	26	27	28		Nist releases dev data		
May	1	2	3	4	5				
	8	9	10	11	12		Nist releases eval data	Sites submit EDT output	
	15	16	17	18	19		NIST returns results		
	22	23	24	25	26		Final Workshop		

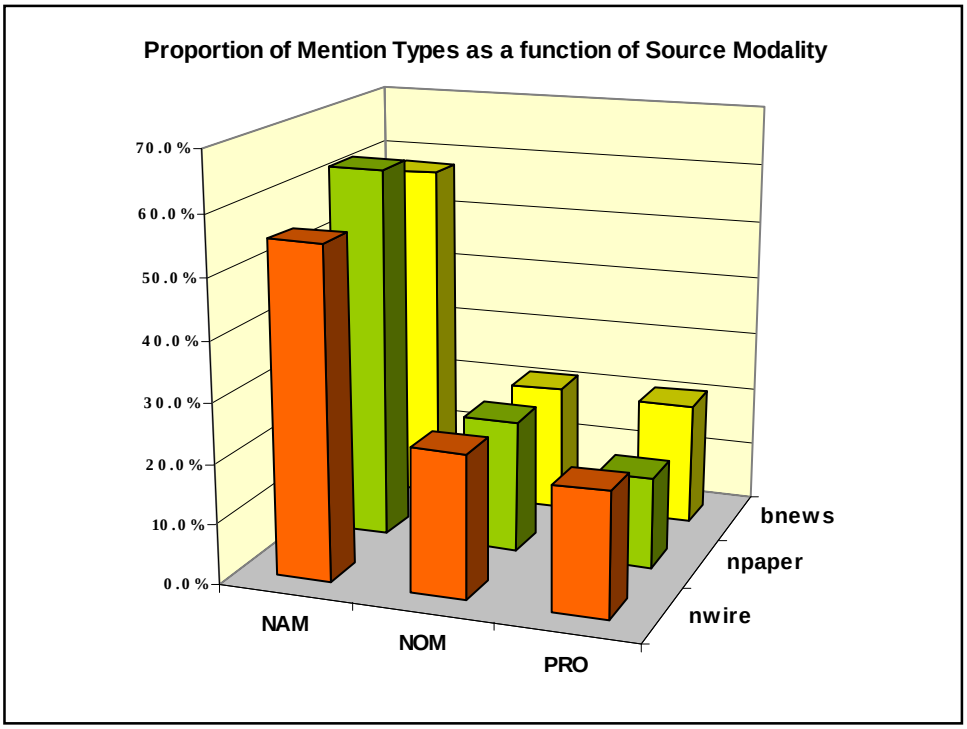
EDT Annotation Assignment for the Pilot Corpus

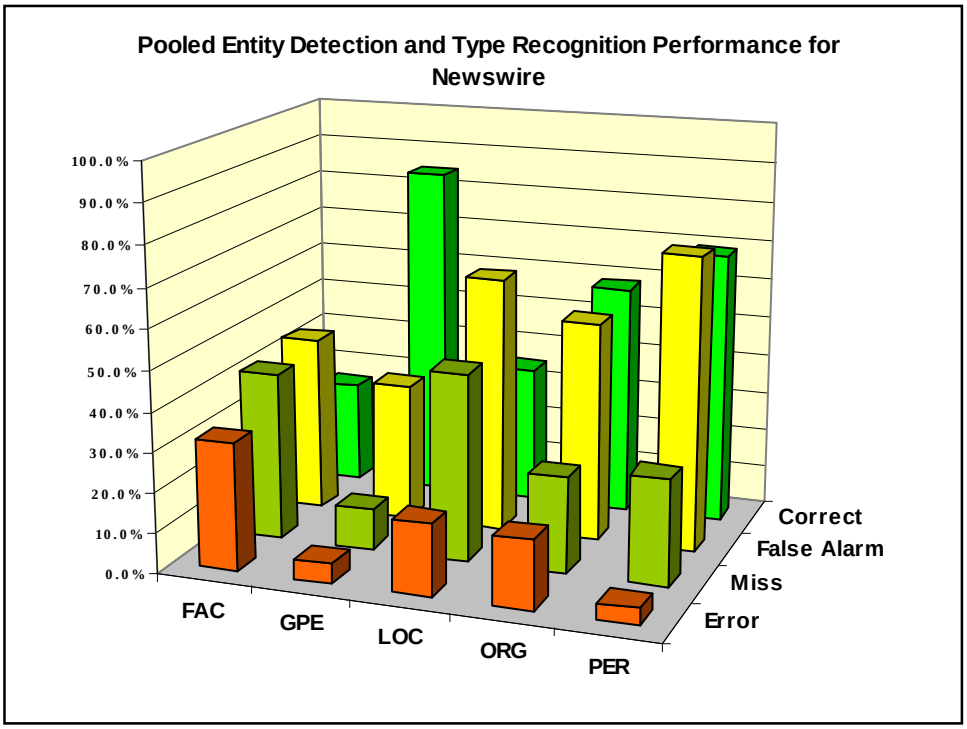
	Training (01-02/98)			Dev Test (03-4/98)			Eval Test (05-06/98)		
Text (newswire)	10,000 words	10,000 words	10,000 words	5,000 words	5,000 words	5,000 words	5,000 words	5,000 words	5,000 words
Audio (broadcast news)	10,000 words	10,000 words	10,000 words	5,000 words	5,000 words	5,000 words	5,000 words	5,000 words	5,000 words
Image (newspaper)	10,000 words	10,000 words	10,000 words	5,000 words	5,000 words	5,000 words	5,000 words	5,000 words	5,000 words
Annotation Team:	**BBN**	**MITRE**	**LDC**	**BBN**	**MITRE**	**LDC**	**BBN**	**MITRE**	**LDC**

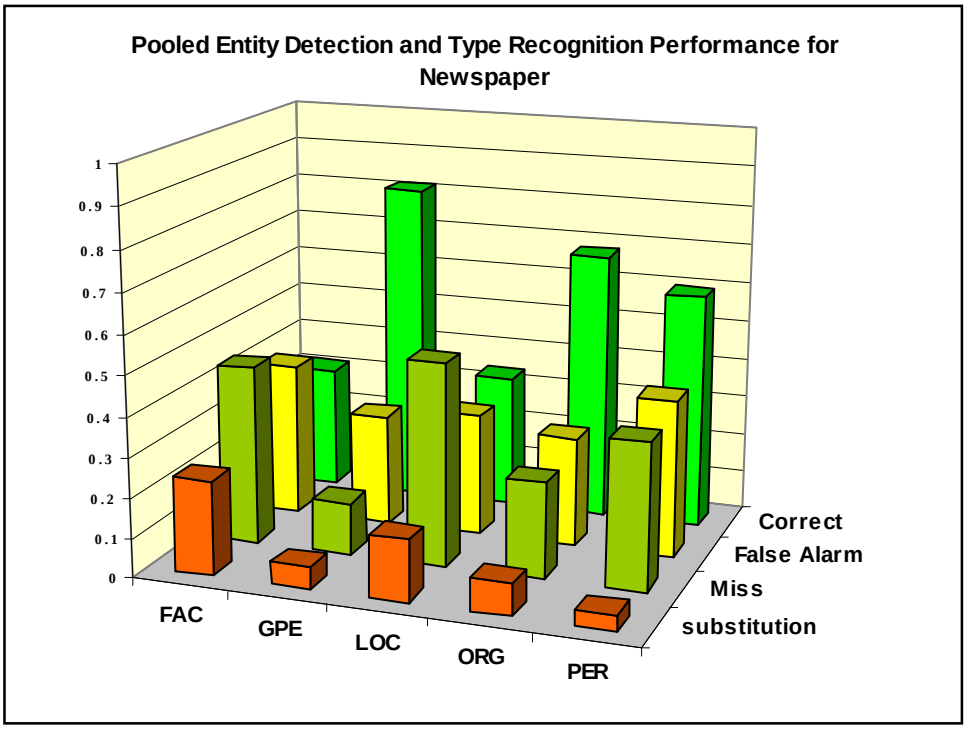


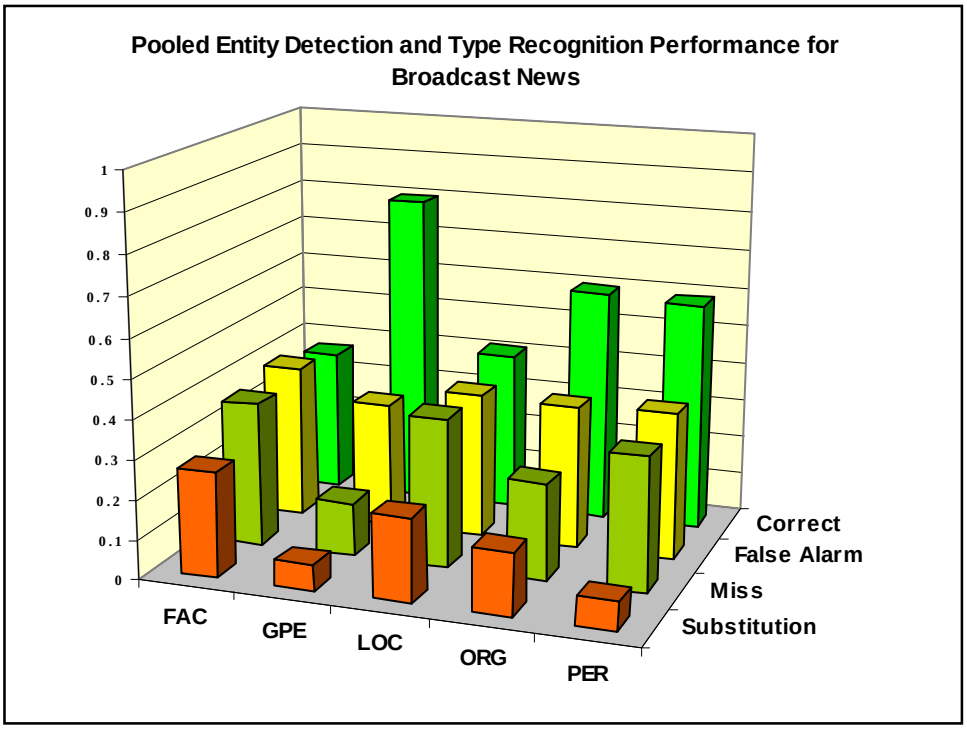

Proportion of Entity Level as a function of Source Modality
70.0 %
60.0 %
50.0 %
40.0 %
30.0 %
20.0 %
10.0 %
0.0 %
NAM
NOM
PRO
bnews
npaper
nwire

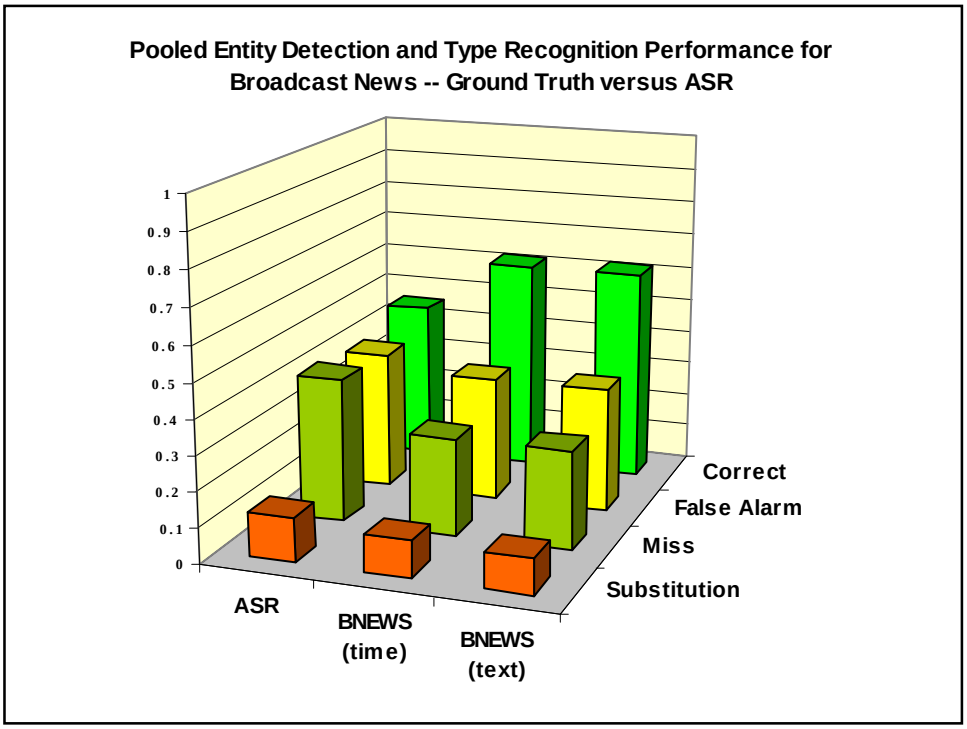

Proportion of Entity Types as a function of Source Modality
70.0 %
60.0 %
50.0 %
40.0 %
30.0 %
20.0 %
10.0 %
0.0 %
FAC
GPE
LOC
ORG
PER
bnews
npaper
nwire

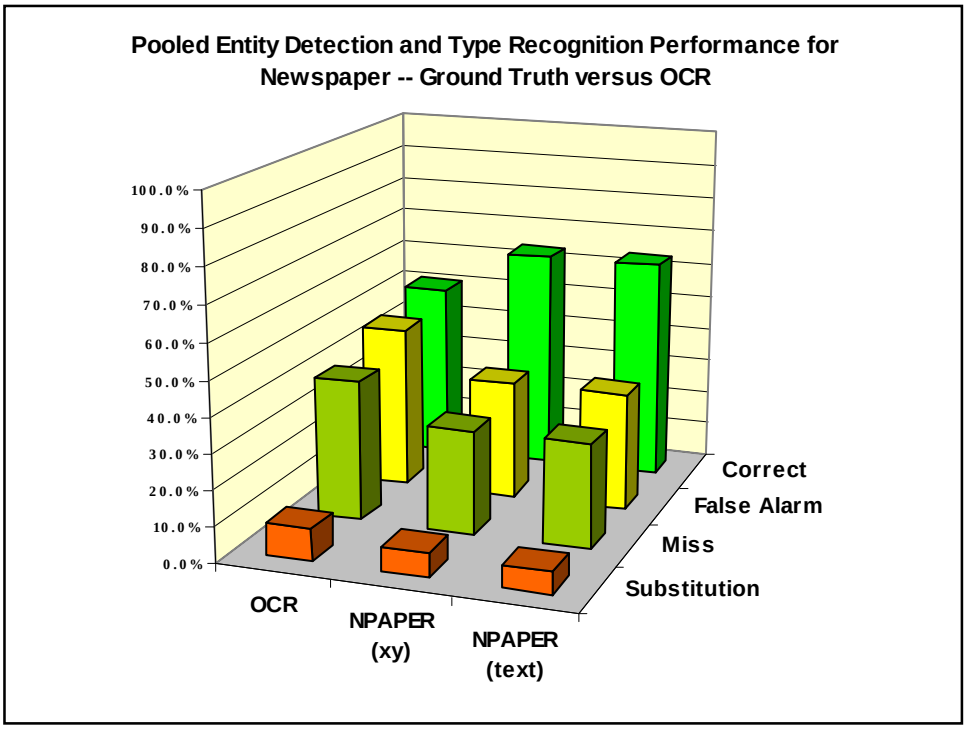

Proportion of Mention Types as a function of Source Modality
70.0%
60.0%
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
NAM
NOM
PRO
bnews
npaper
nwire

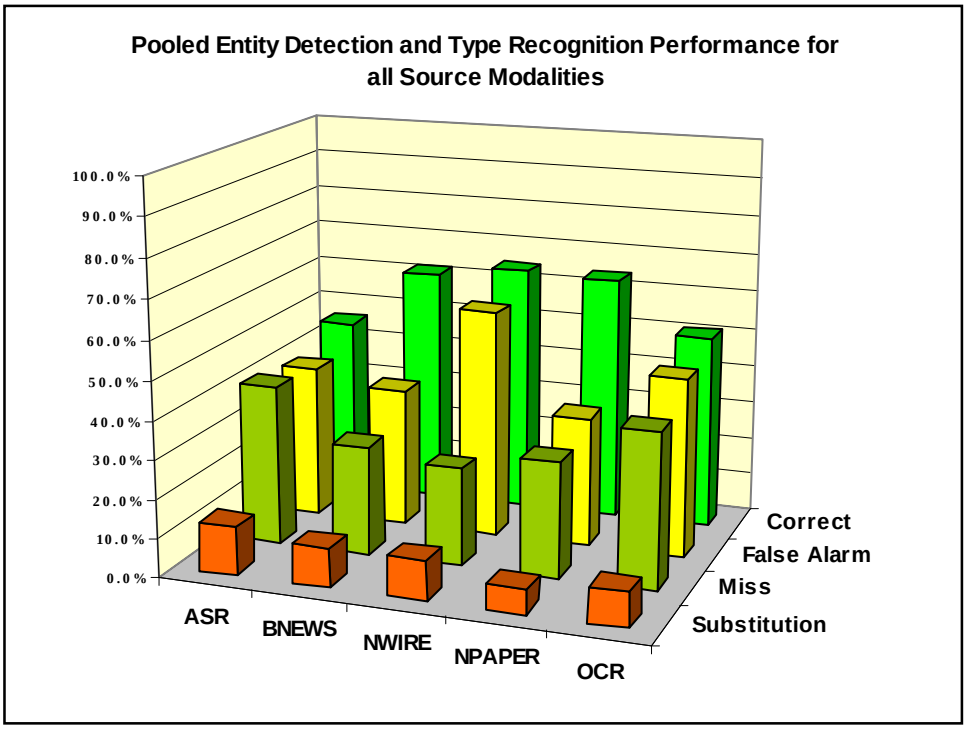

Pooled Entity Detection and Type Recognition Performance for Newswire
100.0%
90.0%
80.0%
70.0%
60.0%
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
FAC
GPE
LOC
ORG
PER
Correct
False Alarm
Miss
Error

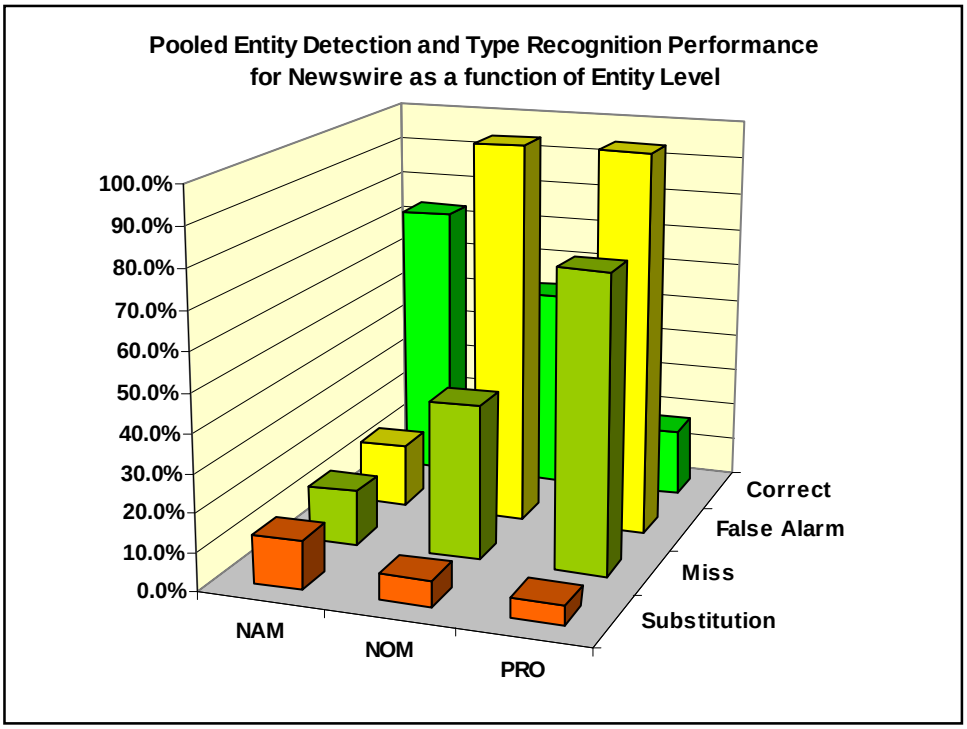

Pooled Entity Detection and Type Recognition Performance for Newspaper
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
FAC
GPE
LOC
ORG
PER
Correct
False Alarm
Miss
substitution

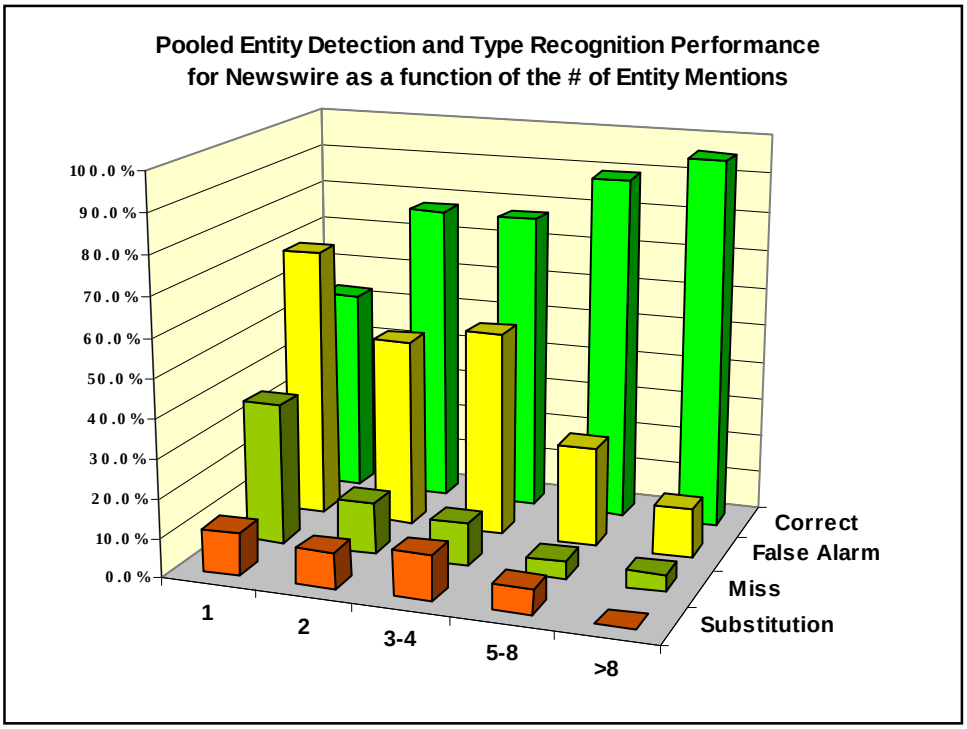

Pooled Entity Detection and Type Recognition Performance for Broadcast News
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
FAC
GPE
LOC
ORG
PER
Correct
False Alarm
Miss
Substitution

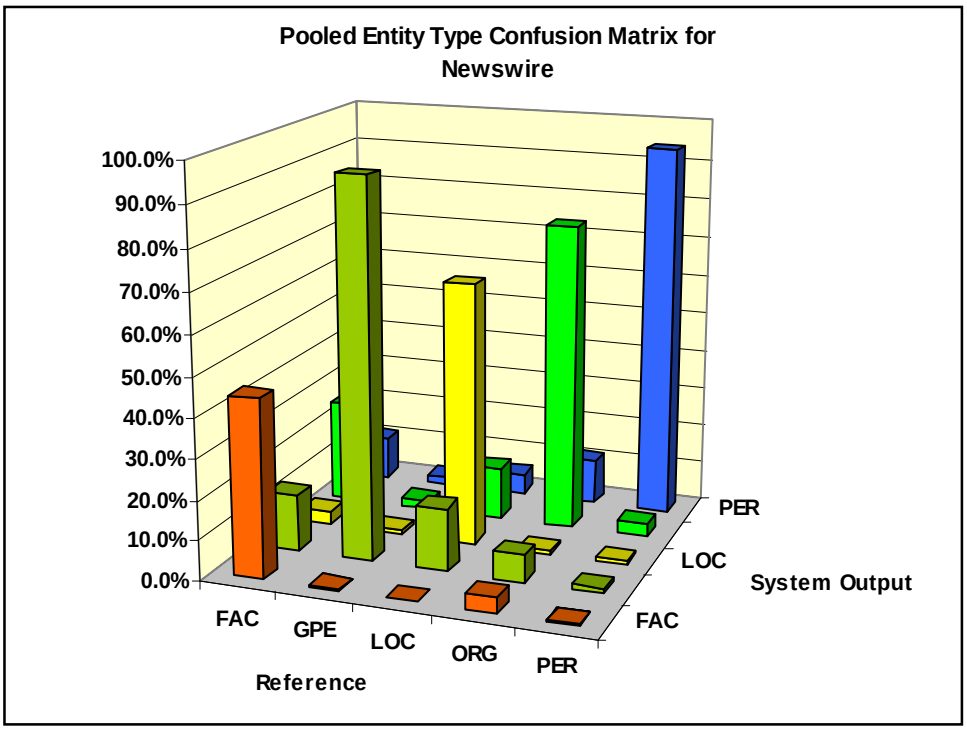

Pooled Entity Detection and Type Recognition Performance for Broadcast News -- Ground Truth versus ASR
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
ASR
BNEWS (time)
BNEWS (text)
Correct
False Alarm
Miss
Substitution

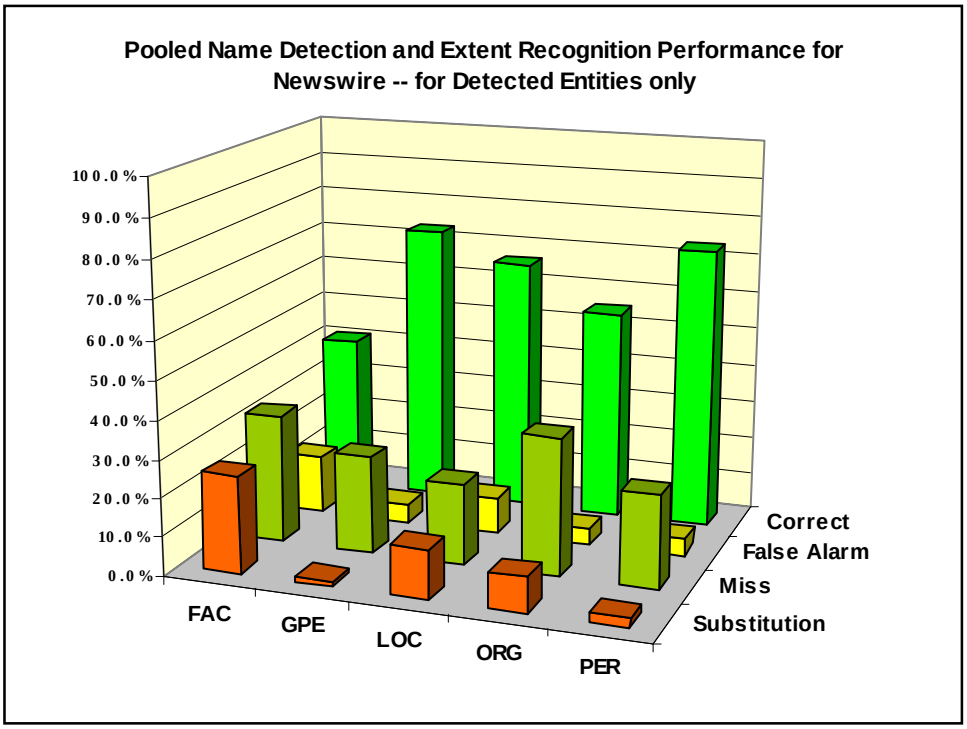

Pooled Entity Detection and Type Recognition Performance for Newspaper -- Ground Truth versus OCR
100.0%
90.0%
80.0%
70.0%
60.0%
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
OCR
NPAPER (xy)
NPAPER (text)
Correct
False Alarm
Miss
Substitution

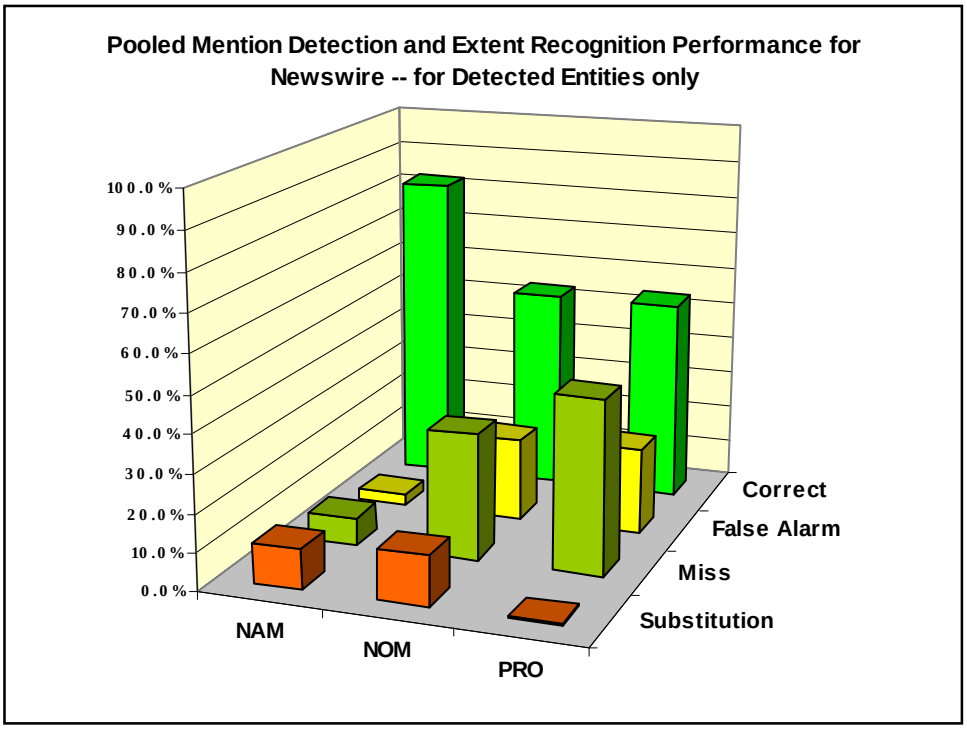

Pooled Entity Detection and Type Recognition Performance for all Source Modalities
100.0%
90.0%
80.0%
70.0%
60.0%
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
ASR
BNEWS
NWIRE
NPAPER
OCR
Correct
False Alarm
Miss
Substitution

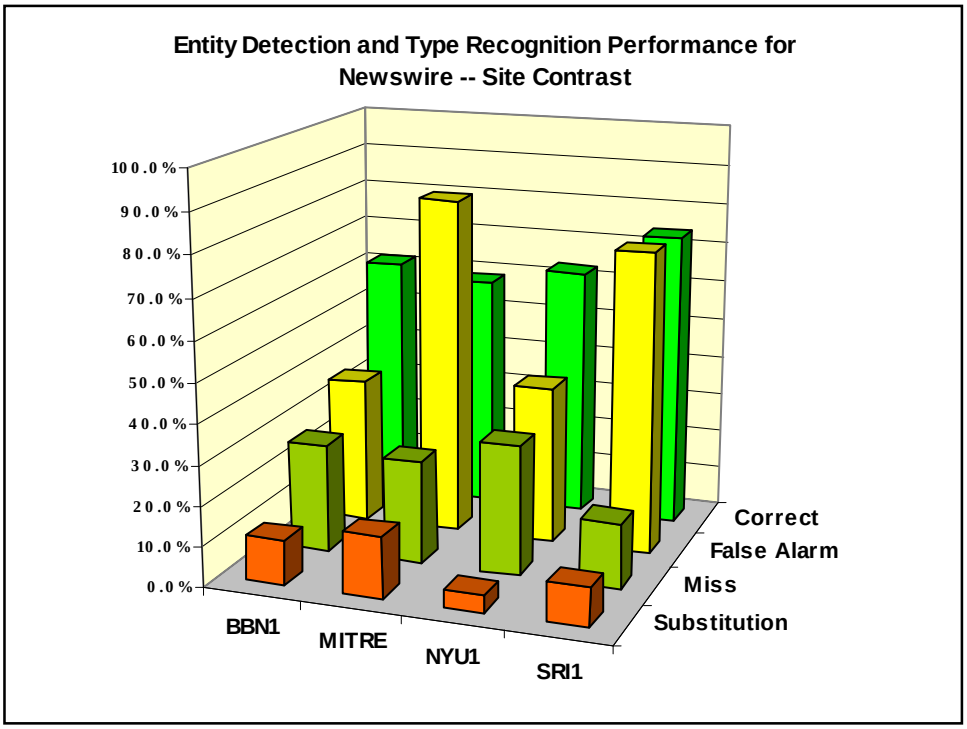

Pooled Entity Detection and Type Recognition Performance for Newswire as a function of Entity Level
100.0%
90.0%
80.0%
70.0%
60.0%
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
NAM
NOM
PRO
Correct
False Alarm
Miss
Substitution

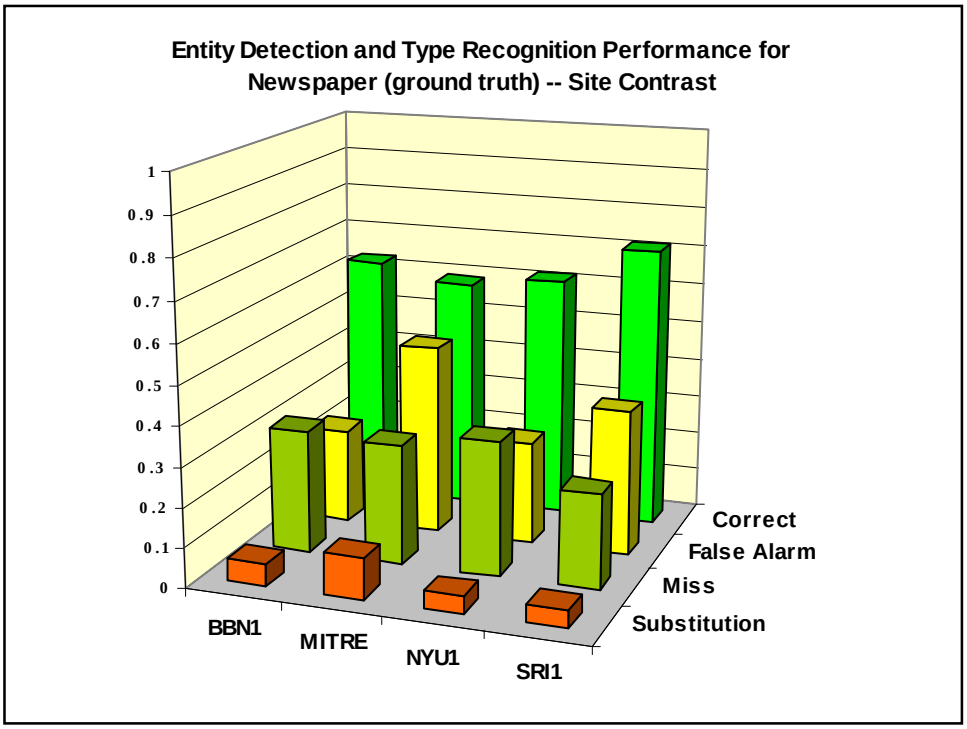

Pooled Entity Detection and Type Recognition Performance for Newswire as a function of the # of Entity Mentions
100.0%
90.0%
80.0%
70.0%
60.0%
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
1
2
3-4
5-8
>8
Correct
False Alarm
Miss
Substitution

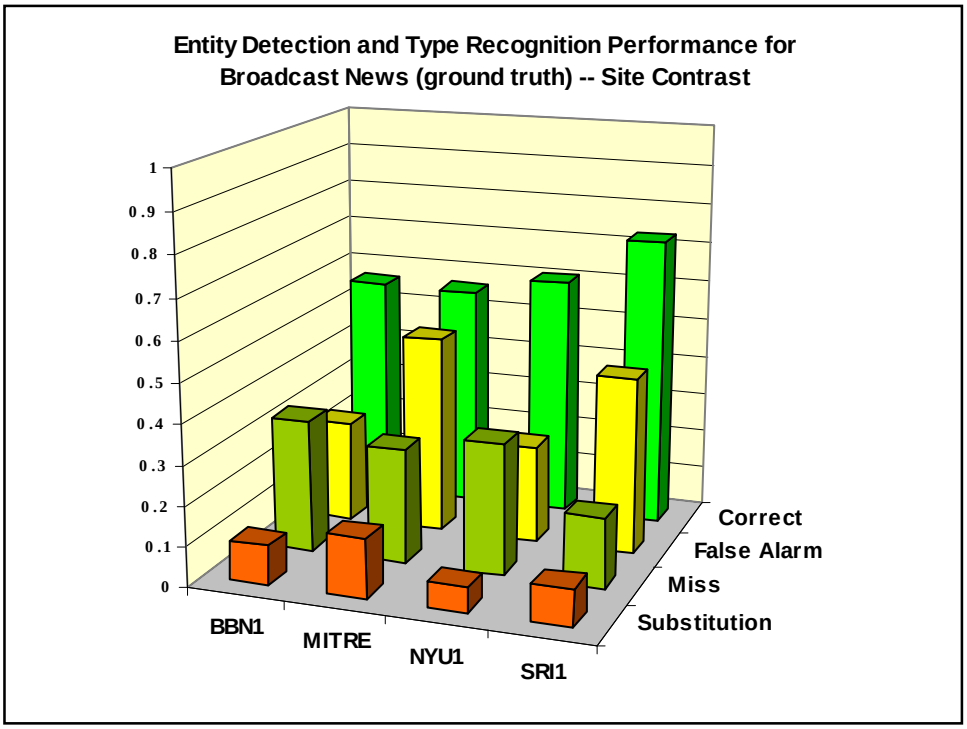

Pooled Entity Type Confusion Matrix for Newswire
100.0%
90.0%
80.0%
70.0%
60.0%
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
FAC
GPE
LOC
ORG
PER
Reference
PER
LOC
FAC
System Output


Pooled Name Detection and Extent Recognition Performance for Newswire -- for Detected Entities only
100.0 %
90.0 %
80.0 %
70.0 %
60.0 %
50.0 %
40.0 %
30.0 %
20.0 %
10.0 %
0.0 %
FAC
GPE
LOC
ORG
PER
Correct
False Alarm
Miss
Substitution


Pooled Mention Detection and Extent Recognition Performance for Newswire -- for Detected Entities only
100.0 %
90.0 %
80.0 %
70.0 %
60.0 %
50.0 %
40.0 %
30.0 %
20.0 %
10.0 %
0.0 %
NAM
NOM
PRO
Correct
False Alarm
Miss
Substitution


Entity Detection and Type Recognition Performance for Newswire -- Site Contrast
100.0%
90.0%
80.0%
70.0%
60.0%
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
BBN1
MITRE
NYU1
SRI1
Correct
False Alarm
Miss
Substitution


Entity Detection and Type Recognition Performance for Newspaper (ground truth) -- Site Contrast
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
BBN1
MITRE
NYU1
SRI1
Correct
False Alarm
Miss
Substitution


Entity Detection and Type Recognition Performance for Broadcast News (ground truth) -- Site Contrast
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
BBN1
MITRE
NYU1
SRI1
Correct
False Alarm
Miss
Substitution

Pilot Study Planning

- Resolve remaining actions, issues and schedule
  - Mark Przybocki will provide ACE sites with sample ASR/OCR source files no later than Monday March 27.
  - David Day will provide working scripts for:
    - converting **ASR/OCR_source** files to **newswire_source** files
    - converting **EDT_newswire_out** files to **EDT_ASR/OCR_out** files

    no later than Monday April 17.

Anything else?…

ACE Program Direction

- Proposed extensions to the EDT task

- Proposed new ACE tasks

Proposed extensions to the EDT task

- New entity types
- New entity attributes
- Role attribute for entity mentions
- Cross-document entity tracking
- Restrict entities to just the important ones
- Restrict mentions to those that are referential
- … <your proposal here>…

New Entity Types

Current	Proposed
• Facility	• FOG (a human-created enterprise = FAC+ORG)
• GSP	• GPE (a geo-political entity = GSP)
• Location	• NGE (a natural geographic entity = LOC)
• Organization	• PER (a person = PER)
• Person	• POS (a place, a spatially determined location)

New Entity Attributes

- ORG: subtype = {government, business, other}
- GPE: subtype = {nation, state, city, other}
- NGE: subtype = {land, water, other}
- PER: nationality = {…}; sex = {M, F, other}
- POS: subtype = {point, line, other}

  Plurality

  dis/conjunctive

Introduce a new concept: The "role" of a mention

- "Entity" is a symbolic construct that represents an abstract identity. Entities have various aspects and functional roles that are associated with their identities.
- We would like to identify these functional roles in addition to identifying the (more abstract) entity identity.
- This may be done by tagging each mention of an entity with its "role", which may be simply one of the (five) "fundamental" entity types.

Proposed new ACE tasks

- Unnumbered tasks
  - Predicate Argument Recognition (aka Proposition Bank)
  - …<your idea here>…
- Numbered tasks
  - …<your idea here>…

Program Planning

- Application ideas
  - Presentations (?)
  - Brainstorming
- Technical infrastructure needs
  - Corpora
  - Tools
- Program direction plans *(Steve Dennis)*

ACE Common Task Candidates (to be evaluated)

- EDT
- Intradoc facts/events (this includes temporal information)
- Xdoc EDT (+ attribute normalization)
- EDT+ (+ = mention roles, more types, metonymy tags, attribute normalization)
- Xdoc facts/events
- Intradoc facts/events+ (+ = modality)
- Predicate Argument Recognition

ACE program activity candidates

- Proposition Bank corpus development
- Create a comprehensive ACE database schema
- Identify a terrific demo for ACE technology