# Scoring Metrics for IDC, CDC and EDC

David Day

ELERFED JHU Workshop

July 18, 2007

# Entity Tracking Tasks

- IDC: Intra-Document Coreference
  - Link all mentions in a document that co-refer to the same entity (out there in the real world)
  - Corpora: MUC, ACE, …
- CDC: Cross-Document Coreference
  - Same as above, but include links that span multiple documents
  - Corpora: John Smith, ACE/ELERFED, (ACE/Culotta?, ACE2003/MITRE, …)
- EDC: Entity Document Categorization
  - For each document D in a set of documents, associate D with all entities that are mentioned at least once within it
  - Corpora: SemEval person biographies, SPOCK
- “Normalization” variants for each of the above
  - Link entity (mentions, documents) to a predefined list of entities

# Metrics That Will be Discussed

- IDC
  - MUC link-based metric (Vilain, et al, 1995)
  - B-CUBED mention-based metric (Baldwin & Bagga)
  - ACE value-based metric (Doddington, et al)
  - Constrained Entity-Alignment F-measure (Luo, 2005)
  - Pairwise Links
  - Edit Distance
- CDC
  - Ditto
- EDC
  - Ditto, plus…
  - F-measure
  - ROC curves?

# Did You Mention Entity? Shifting Terminology

- *mention* (or *entity mention*) $=_{df}$ a phrase referring to an entity in the discourse
  - Earlier authors will sometimes use "entity" to refer to "entity mention" (derived from "named entity expression")
- *entity* (or *equivalence set* of entity mentions, *mention set*, *mention cluster*)
  - Union of all mentions co-referring to the same entity in the world
  - The thing itself (in the real world)

# Desirable Scoring Metric Features

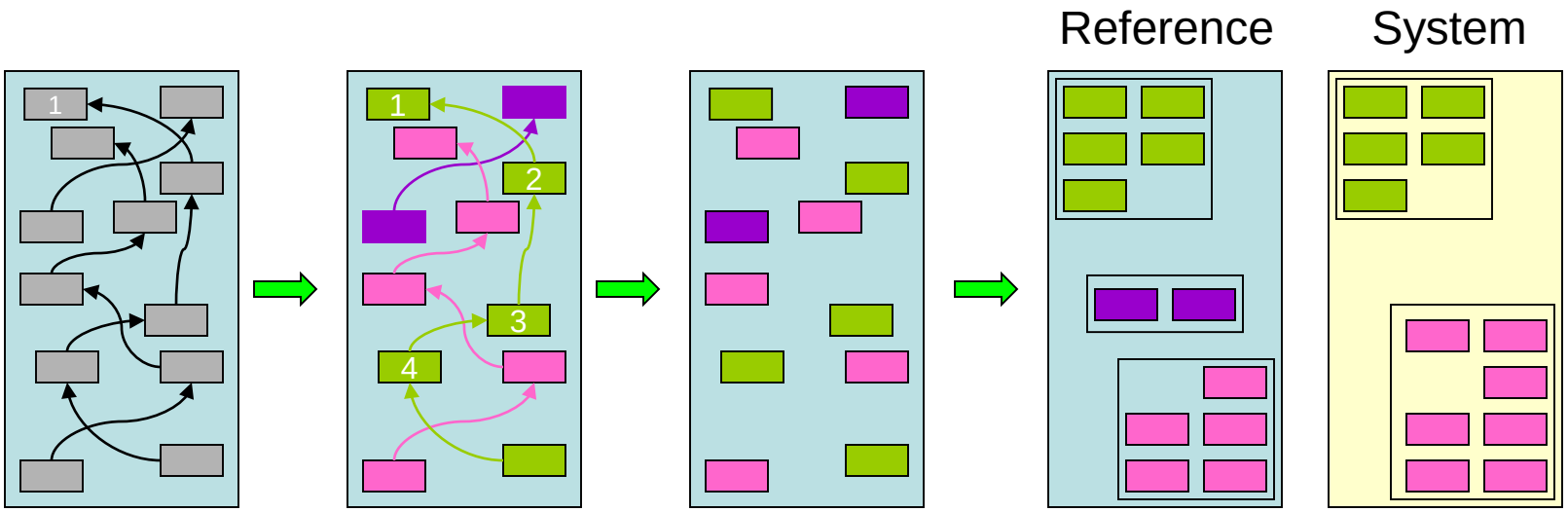
- Discriminability
  - Distinguishes between good and bad system performance levels
  - Ideally, equally across the performance spectrum
- Interpretability
  - Should be relatively “intuitive” to the consumer
- Non-chaotic
  - Small changes in system output should result in relatively small changes in metric value
- Locality?
  - A change in one “part” of system output should not have cascading, non-local effects in the scorer
  - This may be difficult or impossible to achieve, or it may come at the price of some other desirable metric feature
- Community-wide comparability

# MUC-6

- Introduced “model-theory” (sets of mentions) to simplify earlier work that operated directly on link structures
- Involves intersecting ref and sys mention sets, resulting sometimes in non-intuitive scores
  - System output containing a single (completely merged) entity mention set generates 100% recall
  - Identical number of mention sets (entities) can result in identical scores, notwithstanding differing cluster membership
- “Link-based” – measures # of links required to bring sys mention sets into conformance with ref sets
  - Doesn’t account for singleton clusters
  - Undefined for system output containing only singletons
- Intrinsically favors fewer entities
- Tends towards higher scores

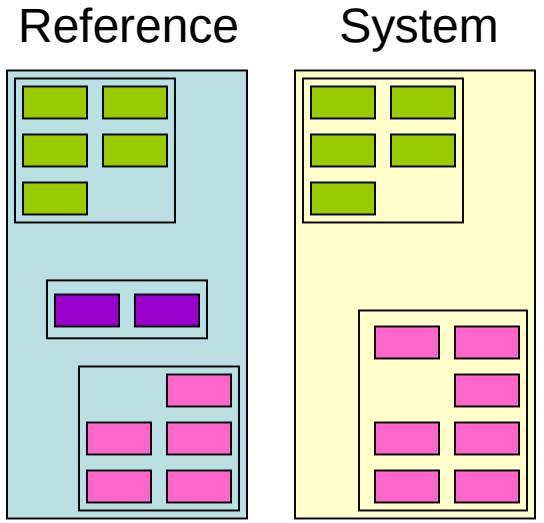
# MUC-6 Co-Reference Scoring Metric (Vilain, et al, 1995)

- Identify the minimum number of link modifications required to make the system mention set identical to the reference set
- Units counted are <u>link edits</u>



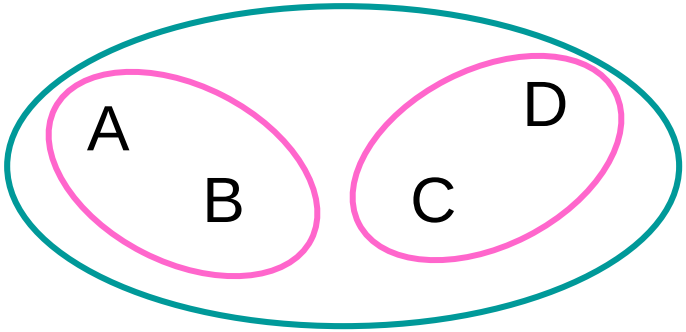
# Model-Theoretic Definitions of Recall and Precision

- S =df Set of key mentions
- p(S) =df Partition of S formed by intersecting all system response sets $R_i$
- Correct links: $c(S) = |S| - 1$
- Missing links: $m(S) = |p(S)| - 1$
- Recall: $\frac{c(S) - m(S)}{c(S)} = \frac{|S| - |p(S)|}{|S| - 1}$
- $\text{Recall}_T = \frac{\sum |S| - |p(S)|}{\sum |S| - 1}$



# MUC-6 Scoring in Action

- Ref = [A, B, C, D]
  Sys = [A, B], [C, D]




Recall $\frac{4 - 2}{3} = 0.66$

Precision $\frac{4 - 1}{4 - 1} = 1.0$

$$\frac{|S| - |p(S)|}{|S| - 1}$$

F-measure $= \frac{2 * 2/3 * 1}{2/3 + 1} = 0.79$

# MUC-6 Scoring
# A More Complicated Example

# B-Cubed

- “Mention-based”
  - Defined for singleton clusters
- Like MUC, relies on intersection operations between ref and sys mention sets
  - Results in sometimes non-intuitive results
  - Tends towards higher scores
    - Entity clusters being used “more than once” within scoring metric is implicated as the likely cause
  - Greater discriminability than the MUC metric
- Incorporates weighting factor
  - CDC setting: equal weighting for each mention (independent of # of mentions in that cluster)
  - “IR” setting: decreases cost of precision errors

# B-Cubed

- Each mention in an equivalence set contributes a fractional amount as a function of the missing mentions

$$\text{Recall} = 1 - \frac{1}{|S|} \sum_j \sum_m \frac{\text{missing}_j(S)}{|S|}$$

$$= 1 - \frac{\sum_j \sum_m |S_i| - |P_{ij}|}{|S|^2}$$

m = mention

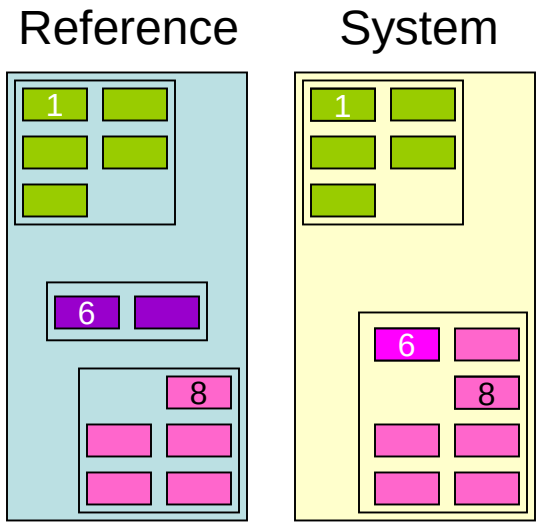
Pij = $j^{th}$ element of the Partition induced on $S_i$ by mentions in system clusters

# B-Cubed Example

$$m_1 = \frac{|\text{mentions}| - |\text{miss}|}{|\text{mentions}|} = \frac{5-0}{5} = 1$$

$$m_6 = \frac{|\text{mentions}| - |\text{miss}|}{|\text{mentions}|} = \frac{7-5}{7} = \frac{2}{7}$$

$$m_8 = \frac{|\text{mentions}| - |\text{miss}|}{|\text{mentions}|} = \frac{7-2}{7} = \frac{5}{7}$$




$$\text{Precision} = \frac{1}{12}\,(m1 + m2 + \ldots\ m6 + m7 + \ldots\ m8\ \ldots)$$

$$= 0.76$$

$$\text{Recall} = 1.0$$

$$\text{F-Measure} = 0.865$$

# Pairwise Links

- Compares entity set membership for each pair of mentions in reference and system
  - If RefM-1 and RefM-2 are in the same cluster, then it is a true-positive if SysM-1 and SysM-2 are as well, and a false-negative otherwise; etc.
  - Simple Recall, Precision and F-measure
- Link-based
  - Not defined for singleton mention clusters
  - Does not rely on MUC, B-Cubed style mention set intersection operations
- Tends towards lower scores than either MUC or B-Cubed
  - Greater discriminability(?)
- Perhaps it's link-based restriction could be fixed without otherwise hurting this metric

# ACE

- Generates one-to-one mapping between ref and sys entities
  - Penalty for un-mapped entities
  - Incorporates dynamic-programming search for mapping that optimizes overall score
  - Mapped entities must share at least one common mention
- EDR Value = $\sum$ sys-token-val / $\sum$ ref-token-val
  - token-val = entity-value * mentions-value
  - Percentage of possible value
  - Can be negative, if too many spurious entities created
- Cost model assigns different weights to entities and mentions
  - Mention type (NAM, NOM, PRO)
  - Entity class (SPC, UPC, GEN)
  - Entity type (PER, GPE, ORG, LOC, FAC, VEH, WEA)
- Difficult to predict how certain system output changes will effect overall score

# ACE Entity and Mentions Value Details

Entity Value

$$Element\_Value(sys) = \begin{cases} \min\left(\prod_{\substack{attribute=\\type,class}} AttrValue(attribute_{sys}), \prod_{\substack{attribute=\\type,class}} AttrValue(attribute_{ref})\right) \cdot \prod_{\substack{attribute=\\type,subtype,class}} W_{err-attribute} & \text{if } sys \text{ mapped} \\ \left(\prod_{\substack{attribute=\\type,class}} AttrValue(attribute_{sys})\right) \cdot W_{FA} & \text{if not mapped} \end{cases}$$

Mentions Value

$$MMV(mention_{sys}) = \begin{cases} \min\left(MTypeValue(mention_{sys}), MTypeValue(mention_{ref})\right) \cdot \prod_{\substack{attribute=\\type,role,style}} W_{Merr-attribute} & \text{if } mention_{sys} \text{ mapped} \\ -MTypeValue(mention_{sys}) \cdot (W_{M-FA} \cdot W_{M-CR}) & \text{if not mapped} \end{cases}$$

# ACE Cost Model

Table 14 Default parameters for scoring EDR performance

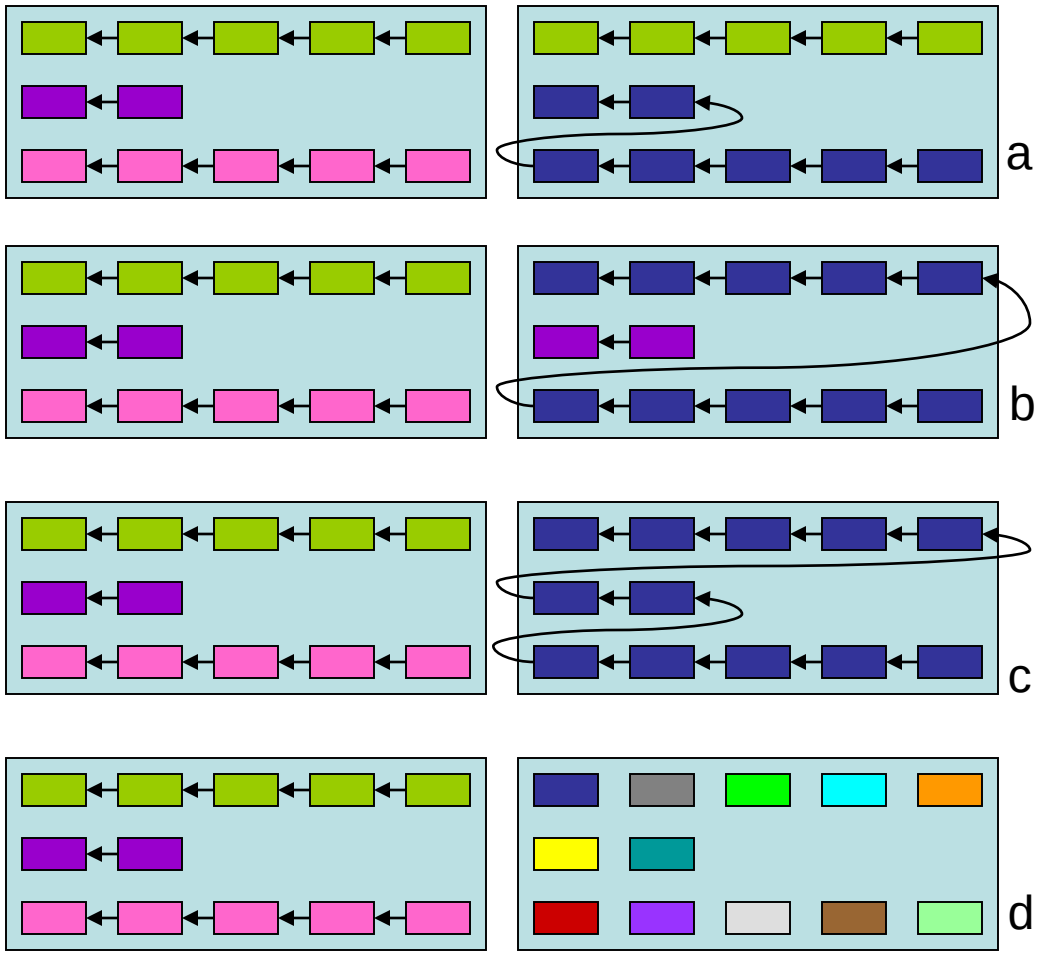
| *Element_Value* parameters | | | |
|---|---|---|---|
| **Attribute** | $W_{err\text{-}attribute}$ | **Attribute Value** | *AttrValue* |
| Type | 0.50 | (all types) | 1.00 |
| Class | 0.75 | SPC | 1.00 |
| | | (not SPC) | 0.00 |
| Subtype | 0.90 | n/a | n/a |
| $W_{E\text{-}FA} = 0.75$ | | | |
| *Mentions_Value*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

# Constrained Entity-Alignment F-Measure Xioqiang Luo, EMNLP 2005

- Like ACE
  - Generates one-to-one mapping
  - Enables independent entity-similarity cost function to be incorporated
  - Search generates optimized score
- Different from ACE
  - Two simpler “entity similarity” cost functions proposed (mention-based vs. entity-based)
    - Mention-based: RefMentions ∩ SysMentions
    - Entity-based: mention F-measure between Ref and Sys
  - Recall and precision computed as a function of ref-to-ref similarity and sys-to-sys similarity, respectively
  - Penalty for over-generation of clusters incorporated directly into precision score
  - Symmetric with respect to Ref and Sys

# Examples

Reference System




| | MUC | B-Cubed | Pair-wise | CEAF (m) | CEAF (e) | ACE |
|---|---|---|---|---|---|---|
| a | 0.947 | 0.865 | | 0.833 | 0.733 | |
| b | 0.947 | 0.737 | | 0.583 | 0.667 | |
| c | 0.900 | 0.545 | 0.40 | 0.417 | 0.294 | |
| d | -------- | 0.400 | ------- | 0.250 | 0.178 | |

| System response | B-cube | | CEAF | | | |
|---|---|---|---|---|---|---|
| | R | P | $\phi_3$-R | $\phi_3$-P | $\phi_4$-R | $\phi_4$-P |
| (c) | 1.0 | 0.375 | 0.417 | 0.417 | 0.196 | 0.588 |
| (d) | 0.25 | 1.0 | 0.250 | 0.250 | 0.444 | 0.111 |

# Comparing CEAF against MUC and ACE on Real Data

| Penalty | #sys-ent | MUC-F | $\phi_3$-CEAF |
|---|---|---|---|
| -0.6 | 561 | .851 | 0.750 |
| -0.8 | 538 | .854 | 0.756 |
| -0.9 | 529 | .853 | 0.753 |
| -1 | 515 | .853 | 0.753 |
| -1.1 | 506 | .856 | 0.764 |
| -1.2 | 483 | .857 | **0.768** |
| -1.4 | 448 | .863 | 0.761 |
| -1.5 | 425 | .862 | 0.749 |
| -1.6 | 411 | .864 | 0.740 |
| -1.7 | 403 | .865 | 0.741 |
| -10 | 113 | **.902** | 0.445 |

| Penalty | #sys-ent | ACE-value(%) | $\phi_3$-CEAF |
|---|---|---|---|
| 0.6 | 1221 | 88.5 | 0.726 |
| 0.4 | 1172 | 89.1 | 0.749 |
| 0.2 | 1145 | 89.4 | 0.755 |
| 0 | 1105 | 89.7 | 0.766 |
| -0.2 | 1050 | **89.7** | 0.775 |
| -0.4 | 1015 | 89.7 | 0.780 |
| -0.6 | 990 | 89.5 | 0.782 |
| -0.8 | 930 | 88.6 | **0.794** |
| -1 | 891 | 86.9 | 0.780 |
| -1.2 | 865 | 86.7 | 0.778 |
| -1.4 | 834 | 85.6 | 0.769 |
| -1.6 | 790 | 83.8 | 0.761 |

# CDC: Entity Detection & Recognition vs. Entity Normalization

- Entity Normalization enables straightforward Recall, Precision and F-measure to be computed trivially
  - No requirements for mapping
  - No need to weight contribution of mentions
  - May want to discount document-level mentions vs. document-level entities

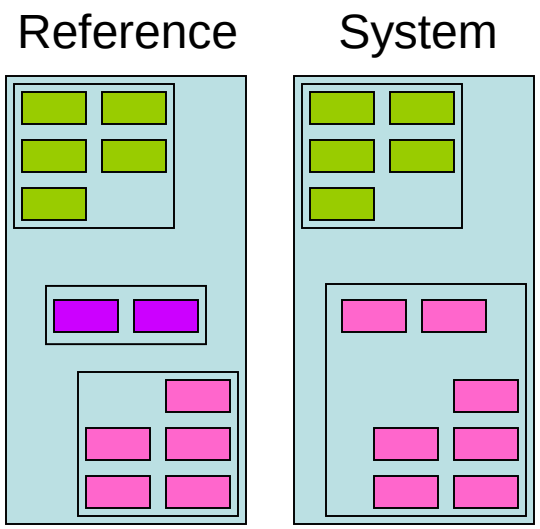
# Some Considerations

- Comparability to community performance measures – MUC, ACE
- Intrinsic scoring metric features
  - Simple, easily interpreted: Pairwise, B-cubed
  - Richly detailed scoring reports: ACE
- Engineering issues
  - Computational costs?
  - (Re-)implementation costs for workshop?
- Optimizing scoring metrics
  - Do these hide “decisions” being made by a system far more powerful than putative end users?

# Baseline Scores

| | IDC-ACE Pub | IDC-ACE ELERFED | CDC-ACE ELERFED | EDC-SemEval Pub | EDC SemEval ELERFED | EDC ACE ELERFED |
|---|---|---|---|---|---|---|
| MUC | | | | | | |
| B-Cubed | | | | | | |
| Pairwise | | | | | | |
| ACE Value | | | | | | |
| CEAF | | | | | | |

# Detritus


Reference
System