

Documentation of ace-eval output

Scoring output

The primary output from ace-eval consists of performance scores for entity (EDR), relation (RDR) and event (VDR) output from ACE systems. These scores are computed and displayed as a function of various conditions, but the output is always in the form shown below, where each record contains:

1. The condition for which the performance statistics are being tabulated
2. The simple count of reference and system objects for this condition, specifically:
 - The number of reference (correct) objects (e.g., the number of reference entities, relations, events, or timex2's)
 - The number of spurious (false alarm) objects that were output by the system
 - The number of reference objects not output (missed) by the system
 - The number of system output objects that were incorrectly recognized.¹
3. The percentage count of reference and system objects (as a fraction of the number of reference objects) for this condition, specifically:
 - The percent of spurious (false alarm) objects that were output by the system
 - The percent of reference objects not output (missed) by the system
 - The percent of system output objects that were incorrect.
 - Precision² = # correct / # output
 - Recall² = # correct / # reference
 - F-measure = 2·Precision·Recall / (Precision + Recall)
4. The cost statistics, in terms of a percentage of the value for perfect model output for this condition, specifically:
 - The cost of spurious (false alarm) output by the system
 - The cost (lost value) of reference objects not output (missed) by the system
 - The cost (lost value) of system output objects that were incorrect

¹ In order to be considered correct:

- An entity must have the correct TYPE, SUBTYPE, and CLASS
- A relation must have the correct TYPE, SUBTYPE, and ARGUMENTS (i.e., the correct argument entities)
- An event must have the correct TYPE, MODALITY, and PARTICIPANTS (i.e., the correct participant entities and the correct roles)
- A timex2 object must have all attributes correctly recognized (i.e., the correct ANCHOR_DIR, ANCHOR_VAL, MOD, SET and VAL)

² Precision and Recall were formulated differently in previous versions of ace-eval, specifically by subtracting only half (rather than all) of the incorrectly recognized objects when tallying the number of objects correctly recognized.

- The cost (lost value) of system output objects that were correct³
- The value of system output = 100% - costFA - costMiss - costError - costCorrect
- Value-based Precision⁴, Pv = $\max(100\% - \text{costFA} - \text{costMiss} - \text{costError} - \text{costCorrect}, 0) / (100\% + \text{costFA})$
- Value-based Recall⁴, Rv = $\max(100\% - \text{costMiss} - \text{costError} - \text{costCorrect}, 0)$
- Value-based F-measure, Fv = $2 \cdot Pv \cdot Rv / (Pv + Rv)$
- 5. Unconditioned cost statistics for this condition, in terms of a percentage of the total value for perfect model output⁵, specifically:
 - The maximum possible value for this condition
 - The cost of spurious (false alarm) output by the system for this condition
 - The cost (lost value) of reference objects not output (missed) by the system for this condition
 - The cost (lost value) of system output objects that were incorrect for this condition
 - The cost (lost value) of system output objects that were correct for this condition

The primary output described above is provided for a number of different conditions for each task. There is also the capability of providing doubly conditioned scores (for example, EDR scores for entities as a function of both type and value). Currently⁶, the conditions are as follows:

EDR scores are shown as a function of the following 6 conditions: ENTITY TYPE, ENTITY LEVEL, ENTITY VALUE, MENTION COUNT, ENTITY CLASS, and SOURCE TYPE.

RDR scores are shown as a function of the following 3 conditions: RELATION TYPE, SOURCE TYPE, and ARGUMENT ERRORS.

VDR scores are shown as a function of the following 4 conditions: EVENT TYPE, EVENT MODALITY, SOURCE TYPE, and PARTICIPANT ERRORS.

Example output for EDR scoring and RDR scoring follows:

³ Objects that are deemed to be “correct” are not necessarily flawless and therefore can suffer a loss of value, according to the cost model being used. For example, an entity that is “correctly” recognized is often not perfectly specified in terms of its mentions, and any missed or spurious mentions of an entity will usually result in a loss of value for that entity. Refer to the cost models for details.

⁴ Values may be negative. Therefore it is necessary to limit value-based Precision and Recall to be non-negative numbers.

⁵ Unconditioned cost statistics are given in order to allow additional supplemental error analyses that aren't provided directly by ace-eval.

⁶ The current version of ace-eval is ace04-eval-v12.

ref	Count				Count_(%)						Cost_(%)				Unconditioned_Cost_(%)								
relation	Rel	Detection	ATSt		Detection	ATSt	Unweighted		Detection	ATSt	ATSt	Value	Value-based	Max	Detection	ATSt	ATSt						
type	Tot	FA	Miss	Err	FA	Miss	Err	Pre--Rec--F	FA	Miss	Err	Corr	(%)	Pre--Rec--F	Value	FA	Miss	Err	Corr				
ART	122	1	83	29	0.8	68.0	23.8	25.0	8.2	12.3	0.1	52.6	18.4	1.4	27.6	58.3	27.6	37.5	5.68	0.00	2.99	1.04	0.08
DISC	150	20	79	59	13.3	52.7	39.3	13.2	8.0	10.0	6.5	30.1	40.7	3.2	19.5	34.1	26.0	29.5	3.45	0.22	1.04	1.40	0.11
EMP-ORG	713	83	157	300	11.6	22.0	42.1	40.1	35.9	37.9	3.9	14.6	20.1	3.9	57.6	68.8	61.5	64.9	41.03	1.60	6.00	8.23	1.59
GPE-AFF	257	27	69	105	10.5	26.8	40.9	38.6	32.3	35.2	5.2	18.4	36.3	5.7	34.4	45.6	39.7	42.4	9.52	0.50	1.75	3.45	0.55
METONYM	9	0	8	1	0.0	88.9	11.1	0.0	0.0	0.0	0.0	77.7	21.5	0.0	0.8	3.7	0.8	1.3	0.15	0.00	0.12	0.03	0.00
OTHER-A	63	3	32	31	4.8	50.8	49.2	0.0	0.0	0.0	1.0	40.6	32.4	0.0	26.0	44.7	27.0	33.7	2.52	0.03	1.02	0.81	0.00
PER-SOC	134	14	27	55	10.4	20.1	41.0	43.0	38.8	40.8	4.1	11.9	28.8	3.0	52.2	61.0	56.3	58.6	18.49	0.76	2.20	5.33	0.55
PHYS	602	71	200	198	11.8	33.2	32.9	43.1	33.9	38.0	6.8	19.6	20.8	3.1	49.7	64.8	56.5	60.4	19.17	1.30	3.76	3.98	0.60
total	2050	219	655	778	10.7	32.0	38.0	38.2	30.1	33.7	4.4	18.9	24.3	3.5	49.0	62.4	53.4	57.5	100.00	4.41	18.88	24.28	3.47

ref	Count				Count_(%)						Cost_(%)				Unconditioned_Cost_(%)								
source	Rel	Detection	ATSt		Detection	ATSt	Unweighted		Detection	ATSt	ATSt	Value	Value-based	Max	Detection	ATSt	ATSt						
type	Tot	FA	Miss	Err	FA	Miss	Err	Pre--Rec--F	FA	Miss	Err	Corr	(%)	Pre--Rec--F	Value	FA	Miss	Err	Corr				
broadca	1036	116	347	381	11.2	33.5	36.8	38.3	29.7	33.5	6.3	24.3	22.9	3.0	43.6	60.8	49.9	54.8	42.04	2.65	10.20	9.62	1.26
newswir	1014	103	308	397	10.2	30.4	39.2	38.2	30.5	33.9	3.0	15.0	25.3	3.8	52.9	63.5	55.9	59.5	57.96	1.76	8.68	14.66	2.21
total	2050	219	655	778	10.7	32.0	38.0	38.2	30.1	33.7	4.4	18.9	24.3	3.5	49.0	62.4	53.4	57.5	100.00	4.41	18.88	24.28	3.47

ref	Count				Count_(%)						Cost_(%)				Unconditioned_Cost_(%)								
argument	Rel	Detection	ATSt		Detection	ATSt	Unweighted		Detection	ATSt	ATSt	Value	Value-based	Max	Detection	ATSt	ATSt						
errors	Tot	FA	Miss	Err	FA	Miss	Err	Pre--Rec--F	FA	Miss	Err	Corr	(%)	Pre--Rec--F	Value	FA	Miss	Err	Corr				
0	1438	219	655	166	15.2	45.5	11.5	61.6	42.9	50.6	7.2	30.7	4.6	5.6	52.0	77.3	59.1	67.0	61.60	4.41	18.88	2.82	3.47
1	500	0	0	500	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	50.4	0.0	49.6	49.6	49.6	33.35	0.00	0.00	16.79	0.00
>1	112	0	0	112	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	92.3	0.0	7.7	7.7	7.7	5.05	0.00	0.00	4.66	0.00
total	2050	219	655	778	10.7	32.0	38.0	38.2	30.1	33.7	4.4	18.9	24.3	3.5	49.0	62.4	53.4	57.5	100.00	4.41	18.88	24.28	3.47