BLANC: Implementing the Rand Index for Coreference Evaluation

Marta Recasens* and Eduard Hovy†

August 14, 2010

1 Motivation

The three motivations behind the Rand index [4], a general clustering evaluation metric, can be rephrased in coreference terms: (i) every mention is unequivocably assigned to a specific entity; (ii) entities are defined just as much by those mentions which they do not contain as by those mentions which they do contain; and (iii) all mentions are of equal importance in the determination of the entity set. The Rand index equals the number of mention pairs that are either placed in an entity in both the **gold partition** (GOLD) and the **system partition** (SYS), or assigned to different entities in both partitions, normalized by the total number of mention pairs. The only use of the Rand index for coreference resolution is by Finkel and Manning [2].

Although Rand has the potential to capture well the coreference problem, it is not useful if applied as originally defined due to the significant imbalance between the number of coreferent mentions and the number of singletons. The extremely high number of mention pairs that are found in different entities in both GOLD and SYS explains the high figures obtained by all systems reported in Finkel and Manning [2], and thus the low discriminatory power of Rand. The measure BLANC (BiLateral Assessment of Noun-phrase Coreference) that we develop implements Rand in a way suited to the coreference problem.

*University of Barcelona, mrecasens@ub.edu
†USC Information Sciences Institute, hovy@isi.edu

2 BLANC: BiLateral Assessment of Noun-phrase Coreference

In order to facilitate future research, we propose BLANC, a measure obtained by applying the Rand index [4] to coreference and taking into account the problems shown by existing measures. BLANC is 'bilateral' in that it takes into consideration both coreference and non-coreference links.

The class-based methods B^3 [1] and CEAF [3] suffer from the essential problem that they reward each link to a class equally no matter how large the class is; assigning a mention to a small class is scored equally as assigning it to a large one. But in principle, assigning it to a large one is making a larger number of pairwise decisions, each of which is equally important. Also, singletons well identified are rewarded like correct full entities. In addition, the MUC metric [5] suffers from the essential problem that it does not explicitly reward correctly identified singletons, yet singletons penalize when included as part of a chain, while it is too lenient with penalizing wrong coreference links.

The Rand index seems to be especially adequate for evaluating coreference since it allows us to measure "non-coreference" as well as coreference links. This makes it possible to address the problem of correctly handling singletons as well as to reward correct coreference chains commensurately with their length. A non-coreference link holds between every two mentions that are deemed to NOT corefer. The interesting property of implementing Rand for coreference is that the sum of all coreference and non-coreference links together is constant for a given set of *N* mentions, namely the triangular number $N(N-1)/2$. By interpreting a system's output as linking each mention to all other mentions as either coreferent or non-coreferent, we can observe the relative distributions within this constant total of coreference and non-coreference links against the gold standard.

In sum then, BLANC models coreference resolution better since it assigns equal importance to every decision of coreferentiality. Further, whereas class-based metrics need to take into account that GOLD and SYS might not contain the same number of entities, and the MUC metric focuses on comparing a possibly unequal number of coreference links, BLANC is grounded on the fact that the total number of links remains constant across GOLD and SYS.

2.1 Coreference and Non-coreference Links

BLANC is best explained considering two kinds of decision:

1. The coreference decisions (made by the coreference system)
   (a) A **coreference link (**c**)** holds between every two mentions that corefer.

(b) A **non-coreference link** (n) holds between every two mentions that do not corefer.

2. The correctness decisions (made by the evaluator)

   (a) A **right link** (r) has the same value (coreference or non-coreference) in GOLD and SYS (i.e., when the system is correct).

   (b) A **wrong link** (w) does not have the same value (coreference or non-coreference) in GOLD and SYS (i.e., when the system is wrong).

Table 1 shows the 2x2 confusion matrix obtained by contrasting the system's coreference decisions against the gold standard decisions. All cells outside the diagonal contain errors of one class being mistaken for the other. BLANC resembles Pairwise F1 as far as coreference links are concerned, but it adds the additional dimension of non-coreference links.

Let *N* be the total number of mentions in a document *d*, and let *L* be the total number of mention pairs (i.e., pairwise links) in *d*, thereby including both coreference and non-coreference links, then

$$L = N(N-1)/2$$

Let *SL* be the total number of links in the SYS partition of *d*, then *SL* equals *L* and is the sum of the four possible types of links:

$$SL = rc + wc + rn + wn$$

where *rc* are the number of right coreference links, *wc* are the number of wrong coreference links, *rn* are the number of right non-coreference links, and *wn* are the number of wrong non-coreference links.

The class imbalance problem of coreference resolution (i.e., the singleton problem) causes that if the Rand index is applied as originally defined by Rand [4], the index concentrates in a small interval near 1 with hardly any discriminatory power. In order to take the under-representation of coreference links into account in the final BLANC score, we compute P, R, and F separately for the two types of link (coreference and non-coreference) and then average them for the final score. The definition of BLANC is shown in Table 2. In BLANC, both coreference and non-coreference links contribute to the final score, but neither more than 50%. BLANC-P and BLANC-R correspond to the average of the two P and R scores, respectively. The final BLANC score corresponds to the average of the two F-scores.

		SYS		Sums
		Coreference	Non-coreference	
GOLD	Coreference	*rc*	*wn*	*rc* + *wn*
	Non-coreference	*wc*	*rn*	*wc* + *rn*
Sums		*rc* + *wc*	*wn* + *rn*	*L*

Table 1: The BLANC confusion matrix.

Score	Coreference	Non-coreference	
P	$P_c = \frac{rc}{rc+wc}$	$P_n = \frac{rn}{rn+wn}$	BLANC-P = $\frac{P_c+P_n}{2}$
R	$R_c = \frac{rc}{rc+wn}$	$R_n = \frac{rn}{rn+wc}$	BLANC-R = $\frac{R_c+R_n}{2}$
F	$F_c = \frac{2P_cR_c}{P_c+R_c}$	$F_n = \frac{2P_nR_n}{P_n+R_n}$	BLANC = $\frac{F_c+F_n}{2}$

Table 2: Definition: Formula for BLANC.

2.2 Boundary Cases

In boundary cases (when for example, SYS or GOLD contain only singletons or only a single set), either P_c or P_n and/or either R_c or R_n are undefined, as one or more denominators will be 0. For these cases we define small variations of the general formula for BLANC, shown in Table 2.

- If SYS contains a single entity, then it only produces coreference links. If GOLD coincides with SYS, BLANC scores equal 100. If GOLD is fully the dual (i.e., it contains only singletons), BLANC scores equal 0. Finally, if GOLD contains links of both types, P_n, R_n and F_n equal 0.

- If SYS contains only singletons, then it only produces non-coreference links. If GOLD coincides with SYS, BLANC scores equal 100. If GOLD is fully the dual (i.e., it contains a single entity), BLANC scores equal 0. Finally, if GOLD contains links of both types, P_c, R_c and F_c equal 0.

- If GOLD includes links of both types but SYS contains no right coreference link, then P_c, R_c and F_c equal 0. Instead, if SYS contains no right non-coreference link, then P_n, R_n and F_n equal 0.

- If SYS contains links of both types but GOLD contains a single entity, BLANC scores equal P_c, R_c and F_c. Instead, if GOLD contains only singletons, BLANC scores equal P_n, R_n and F_n.

A near-boundary case reveals the main weakness of BLANC. This is the case in which all links but one are non-coreferent and the system outputs only non-coreference links. Then, the fact that BLANC places equal importance on the one link as on all the remaining links together leads to a too severe penalization, as the BLANC score will never be higher than 50. One can either simply accept this as a quirk of BLANC or, following the beta parameter used in the F-score, one can introduce a parameter that enables the user to change the relative weights given to coreference and non-coreference links.

References

[1] Bagga, A., and Baldwin, B. 1998. Algorithms for scoring coreference chains. In *Proceedings of the LREC 1998 Workshop on Linguistic Coreference*, pp. 563–566. Granada, Spain.

[2] Finkel, J. R., and Manning, C. D. (2008) Enforcing transitivity in coreference resolution. In *Proceedings of ACL-HLT 2008*, pp. 45–48. Columbus, Ohio.

[3] Luo, X. 2005. On coreference resolution performance metrics. In *Proceedings of HLT-EMNLP 2005*, pp. 25–32. Vancouver, Canada.

[4] Rand, W. M. (1971) Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850.

[5] Vilain, M., Burger, J., Aberdeen, J., Connolly, D., and Hirschman, L. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of MUC-6*, pp. 45–52. San Francisco, California.