Natural Language Engineering **17** (4): 485–510. © Cambridge University Press 2010
doi:10.1017/S135132491000029X

BLANC: Implementing the Rand index for coreference evaluation

M. RECASENS[1] and E. HOVY[2]
[1]*CLiC, University of Barcelona, Gran Via 585, Barcelona 08007, Spain*
email: mrecasens@ub.edu
[2]*USC Information Sciences Institute, 4676 Admiralty Way, Marina del Rey, CA 90292, USA*
email: hovy@isi.edu

(*Received 3 May 2010; revised 17 August 2010; accepted 28 October 2010; first published online 6 December 2010*)

Abstract

This paper addresses the current state of coreference resolution evaluation, in which different measures (notably, MUC, B^3, CEAF, and ACE-value) are applied in different studies. None of them is fully adequate, and their measures are not commensurate. We enumerate the desiderata for a coreference scoring measure, discuss the strong and weak points of the existing measures, and propose the BiLateral Assessment of Noun-Phrase Coreference, a variation of the Rand index created to suit the coreference task. The BiLateral Assessment of Noun-Phrase Coreference rewards both coreference and non-coreference links by averaging the F-scores of the two types, does not ignore singletons – the main problem with the MUC score – and does not inflate the score in their presence – a problem with the B^3 and CEAF scores. In addition, its fine granularity is consistent over the whole range of scores and affords better discrimination between systems.

1 Introduction

Coreference resolution is the task of determining which expressions in a text refer to the same entity or event. At heart, the problem is one of grouping into 'equivalence classes' all mentions that corefer and none that do not, which is a kind of clustering. But since documents usually contain many referring expressions, many different combinations are possible, and measuring partial cluster correctness, especially since *sameness* is transitive, makes evaluation difficult. One has to assign scores to configurations of correct and incorrect links in a way that reflects intuition and is consistent. Different assignment policies have resulted in different evaluation measures that deliver quite different patterns of scores. Among the different scoring measures that have been developed, the following four are generally used: MUC (Vilain *et al.* 1995), B^3 (Bagga and Baldwin 1998), CEAF (Luo 2005), and the Automatic Content Extraction (ACE) value (Doddington *et al.* 2004).

Unfortunately, despite the measures being incommensurate, researchers often use only one or two measures when evaluating their systems. For example, some people employ the (older) MUC measure in order to compare their results with previous

https://doi.org/10.1017/S135132491000029X Published online by Cambridge University Press

[Eyewitnesses]$_{m_1}$ reported that [Palestinians]$_{m_2}$ demonstrated today Sunday in [the West Bank]$_{m_3}$ against [the [Sharm el-Sheikh]$_{m_4}$ summit to be held in [Egypt]$_{m_6}$]$_{m_5}$. In [Ramallah]$_{m_7}$, [around 500 people]$_{m_8}$ took to [[the town]$_{m_9}$'s streets]$_{m_{10}}$ chanting [slogans]$_{m_{11}}$ denouncing [the summit]$_{m_{12}}$ and calling on [Palestinian leader Yasser Arafat]$_{m_{13}}$ not to take part in [it]$_{m_{14}}$.

Fig. 1. Example of coreference (from ACE-2004).

work (Haghighi and Klein 2007; Yang *et al.* 2008); others adopt the more recent advances and use either B^3, CEAF, or the ACE-value (Daumé and Marcu 2005; Culotta *et al.* 2007); and a third group includes two or more scores for the sake of completeness (Luo *et al.* 2004; Bengtson and Roth 2008; Finkel and Manning 2008; Poon and Domingos 2008; Ng 2009).

This situation makes it hard to successfully compare systems, hindering the progress of research in coreference resolution. There is a pressing need to (1) define what exactly a scoring metric for coreference resolution needs to measure; (2) understand the advantages and disadvantages of each of the existing measures; and (3) reach agreement on a standard measure(s). This paper addresses the first two questions – we enumerate the desiderata for an adequate coreference scoring measure, and compare the different existing measures – and proposes the BiLateral Assessment of Noun-Phrase Coreference (BLANC) measure. BLANC adapts the Rand index (Rand 1971) to coreference addressing observed shortcomings in a simple fashion to obtain a fine granularity that allows better discrimination between systems.

The paper is structured as follows. Section 2 considers the difficulties of evaluating coreference resolution. Section 3 gives an overview of the existing measures, highlighting their advantages and drawbacks, and lists some desiderata for an ideal measure. In Section 4, the BLANC measure is presented in detail. Section 5 shows the discriminative power of BLANC by comparing its scores with those of the other measures on artificial and real data, and provides illustrative plots. Finally, conclusions are drawn in Section 6.

2 Coreference resolution and its evaluation: an example

Coreference resolution systems assign each mention (usually a noun phrase) in the text to the entity it refers to and thereby link coreferent mentions into chains.[1] Some entities are expressed only once (singletons), whereas others are referred multiple times (multimention entities). Only multimention entities contain coreferent mentions. For example, in the text segment of Figure 1, we find the following:

- Nine singletons: {*eyewitnesses*}$_{G1}$, {*Palestinians*}$_{G2}$, {*the West Bank*}$_{G3}$, {*Sharm el-Sheikh*}$_{G4}$, {*Egypt*}$_{G5}$, {*around 500 people*}$_{G6}$, {*the town's streets*}$_{G7}$, {*slogans*}$_{G8}$, {*Palestinian leader Yasser Arafat*}$_{G9}$.
- One two-mention entity: {*Ramallah*, *the town*}$_{G10}$.

[1] Following the terminology of the ACE program, a *mention* is defined as an instance of reference to an object, and an *entity* is the collection of mentions referring to the same object in a document.

https://doi.org/10.1017/S135132491000029X Published online by Cambridge University Press

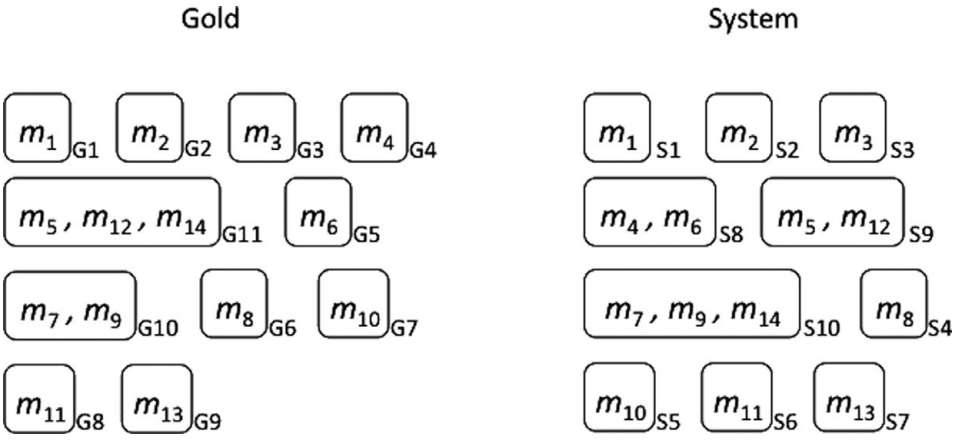


Fig. 2. The problem of comparing the gold partition with the system partition for a given text (Figure 1).

- One three-mention entity: {*the Sharm el-Sheikh summit to be held in Egypt*, *the summit*, *it*}$_{G11}$.

In evaluating the output produced by a coreference resolution system, we need to compare the true set of entities (the *gold partition*, GOLD, produced by human expert) with the predicted set of entities (the *system partition*, SYS, produced by the system or human to be evaluated). The mentions in GOLD are known as true mentions, and the mentions in SYS are known as system mentions. Let a system produce the following partition for the same example given in Figure 1:

- Seven singletons: {*eyewitnesses*}$_{S1}$, {*Palestinians*}$_{S2}$, {*the West Bank*}$_{S3}$, {*around 500 people*}$_{S4}$, {*the town's streets*}$_{S5}$, {*slogans*}$_{S6}$, {*Palestinian leader Yasser Arafat*}$_{S7}$.
- Two two-mention entities: {*Sharm el-Sheikh*, *Egypt*}$_{S8}$, {*the Sharm el-Sheikh summit to be held in Egypt*, *the summit*}$_{S9}$.
- One three-mention entity: {*Ramallah*, *the town*, *it*}$_{S10}$.

Schematically, the comparison problem is illustrated in Figure 2. Some links are missed and others are wrongly predicted; e.g. entity S9 is missing one mention (compare with G11), whereas S10 includes a wrong mention, and two non-coreferent mentions are linked under S8. The difficulty of evaluating coreference resolution arises from the interaction of the issues that have to be addressed simultaneously: Should we focus on the number of correct coreference links? Or should we instead take each equivalence class as the unit of evaluation? Do we reward singletons with the same weight that we reward a multimention entity? Different decisions will result in different evaluation scores, which will determine how good SYS is considered to be in comparison with GOLD.

The evaluation measures developed to date all make somewhat different decisions on these points. While these decisions have been motivated in terms of one or another criterion, they also have unintended and unsatisfactory consequences. We

https://doi.org/10.1017/S135132491000029X Published online by Cambridge University Press

Table 1. *Comparison of evaluation metrics on the examples in Figure 3*

System response	MUC-F	B^3-F	CEAF	F1	H	Rand
(a)	94.7	86.5	83.3	80.8	77.8	84.8
(b)	94.7	73.7	58.3	63.6	57.1	62.1
(c)	90.0	54.5	41.7	48.3	0	31.8
(d)	–	40.0	25.0	–	48.7	68.2

next review some current measures and identify the desiderata for a coreference measure.

3 Current measures and desiderata for the future

3.1 Current measures: strong and weak points

This section reviews the main advantages and drawbacks of the principal coreference evaluation measures. The main difference resides in the way they conceptualize how a coreference set within a text is defined: either in terms of *links*, i.e. the pairwise links between mentions (MUC, Pairwise F1, Rand), or in terms of *classes* or *clusters*, i.e. the entities (B^3, CEAF, ACE-value, mutual information). Although the two approaches are equivalent in that knowing the links allows building the coreference classes, and knowing the classes allows inferring the links, differences in instantiation design produce a range of evaluation metrics that vary to such an extent that still today there is no widely agreed upon standard. Table 1 shows how the different system outputs in Figure 3 (borrowed from Luo (2005)) are scored by the various scoring algorithms presented next.

MUC (Vilain *et al.* 1995). This is the oldest and most widely used measure, defined as part of the MUC-6 and MUC-7 evaluation tasks on coreference resolution. It relies on the notion that the minimum number of links needed to specify either GOLD or SYS is the total number of mentions minus the number of entities. The MUC measure computes the number of all coreference links common between GOLD and SYS. To obtain recall (R), this number is divided by the minimum number of links required to specify GOLD. To obtain precision (P), it is divided by the minimum number of links required to specify SYS.

As observed by Bagga and Baldwin (1998) and Luo (2005), the MUC metric is severely flawed for two main reasons. First, it is indulgent, as it is based on the *minimal* number of missing and wrong links, which often results in counterintuitive results. Classifying one mention into a wrong entity counts as one P and one R error, while completely merging two entities counts as a single R error, although this is further away from the real answer. As a result, the MUC score is too lenient with systems that produce overmerged entities (entity sets containing many referring expressions), as shown by system responses (b) and (c) in Table 1. If all mentions

https://doi.org/10.1017/S135132491000029X Published online by Cambridge University Press

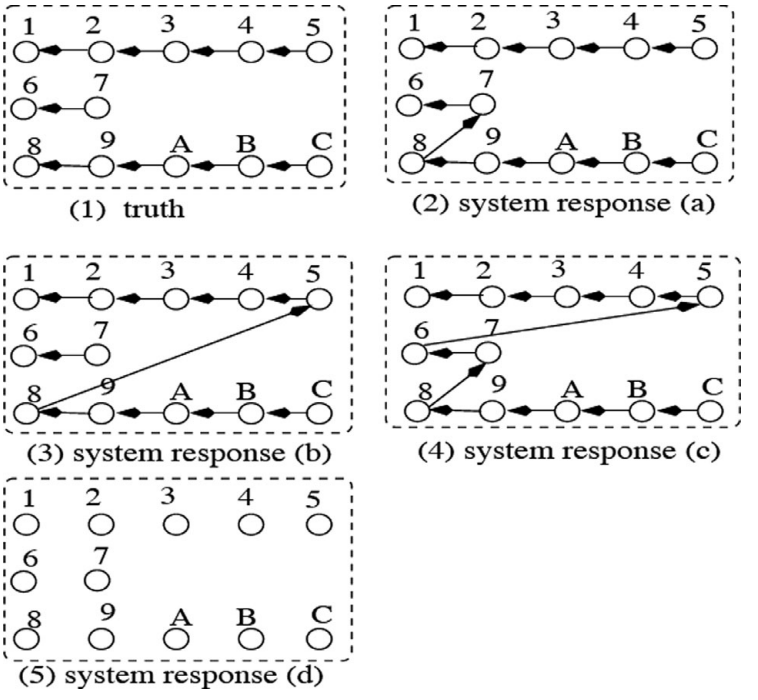


Fig. 3. Example entity partitions (from Luo (2005)).

in each document of the MUC test sets[2] are linked into one single entity, the MUC metric gives a score higher than any published system (Finkel and Manning 2008). Second, given that it only takes into account coreference links, the addition of singletons to SYS does not make any difference. It is only when a singleton mention is misclassified in a multimention entity that there is decrease in the MUC score. This is why the entry for system response (d) in Table 1 is empty.

B^3 (Bagga and Baldwin 1998). To penalize clustering too many mentions in the same entity, this metric computes R and P for each mention, including singletons. The total number of intersecting mentions between the GOLD and the SYS entities is computed and divided by the total number of mentions in the GOLD entity to obtain R, or in the SYS entity to obtain P. The average over the individual mention scores gives the final scores.

Although B^3 addresses the shortcomings of MUC, it presents a drawback in that scores squeeze up too high due to singletons: when many singletons are present, scores rapidly approach 100 per cent. This leaves little numerical room for comparing systems, and forces one to consider differences in the second and third decimal places when scores are high (while such differences are meaninglessly small in lower ranges). It is not possible to observe this in Table 1, as the truth in Figure 3 does not contain

[2] The MUC-6 and MUC-7 corpora were only annotated with multimention entities (Hirschman and Chinchor 1997).

https://doi.org/10.1017/S135132491000029X Published online by Cambridge University Press

Table 2. *Distribution of mentions into entities in two corpora: the English ACE-2004 (Doddington et al. 2004) and the Spanish AnCora-Es (Recasens and Martí 2010)*

	ACE-2004 (English)		AnCora-Es (Spanish)	
	(#)	(%)	(#)	(%)
Mentions	28,880	100.00	88,875	100.00
Entities	11,989	100.00	64,421	100.00
Singletons	7,305	60.93	55,264	85.79
2-mention	2,126	17.73	4,825	7.49
3-mention	858	7.16	1,711	2.66
4-mention	479	4.00	869	1.35
5-mention	287	2.39	485	0.75
6–10-mention	567	4.73	903	1.40
>11-mention	367	3.06	364	0.57

any singleton. However, it turns out that singletons are the largest group in real texts (see Table 2): about 86 per cent of the entities if the entire set of mentions is considered, like in the AnCora corpora; 61 per cent of the entities in the ACE corpora, where the coreference annotation is restricted to seven semantic types (person, organization, geo-political entity, location, facility, vehicle, and weapon). A side effect is that B^3 scores are inflated, obscuring the intuitively appropriate level of accuracy of a system in terms of coreference links.

CEAF (Luo 2005). Luo (2005) considers that B^3 can give counterintuitive results due to the fact that an entity can be used more than once when aligning the entities in GOLD and SYS. In Figure 3, B^3–R is 100 per cent for system response (c) even though the true set of entities has not been found; conversely, B^3–P is 100 per cent for system response (d) even though not all the SYS entities are correct. Thus, he proposes CEAF, which finds the best one-to-one mapping between the entities in GOLD and SYS, i.e. each SYS entity is aligned with at most one GOLD entity, and the best alignment is the one maximizing the similarity. Depending upon the similarity function, Luo (2005) distinguishes between the mention-based CEAF and the entity-based CEAF, but we will focus on the former as it is the most widely used. It employs Luo's (2005) ϕ_3 similarity function. When true mentions are used, R and P scores are the same. They correspond to the number of common mentions between every two aligned entities divided by the total number of mentions.

CEAF, however, suffers from the singleton problem just as B^3 does. This accounts for the fact that the B^3 and CEAF scores are usually higher than the MUC on corpora where singletons are annotated (e.g. ACE, AnCora), because a great percentage of the score is simply due to the resolution of singletons. In addition, CEAF's entity alignment might cause a correct coreference link to be ignored if that entity finds no alignment in GOLD (Denis and Baldridge 2009). Finally, all entities are weighted equally, irrespective of the number of mentions they contain (Stoyanov *et al.* 2009), so that creating a wrong entity composed of two small entities

https://doi.org/10.1017/S135132491000029X Published online by Cambridge University Press

is penalized to the same degree as creating a wrong entity composed of a small and a large entity.

ACE-value (Doddington *et al.* 2004). The ACE-value, the official metric in the ACE program, is very task-specific, and not really useful for the general coreference problem that is not limited to a set of specific semantic types. A score is computed by subtracting a normalized cost from 1. The normalized cost corresponds to the sum of errors produced by unmapped and missing mentions/entities as well as wrong mentions/entities,[3] normalized against the cost of a system that does not output any entity. Each error has an associated cost that depends on the type of the ACE-entity and on the kind of mention, but these costs have changed between successive evaluations. The ACE-value is hard to interpret (Luo 2005): a system with 90 per cent does not mean that 90 per cent of system entities or mentions are correct, but that the cost of the system, relative to the one producing no entity, is 10 per cent.

Pairwise F1. Also known as positive-link-identification F-score. If reported, this metric is always included in addition to MUC, B^3 and/or CEAF, as it is meant to give some further insight not provided by the other metrics (Choi and Cardie 2007; Poon and Domingos 2008; Haghighi and Klein 2009). Pairwise F1 simply computes P, R, and F over all pairs of coreferent mentions. As noted by Haghighi and Klein (2009), merging or separating entities is over-penalized quadratically in the number of mentions. Besides, it ignores the correct identification of singletons.

Mutual information, H (Popescu-Belis 2000). The H measure draws on information theory to evaluate coreference resolution. GOLD and SYS are seen as the two ends of the communication channel, GOLD being the sender or speaker, and SYS being the receiver or the hearer. The coreference information of GOLD and SYS correspond to the entropy of GOLD and SYS, respectively. Then the GOLD and SYS partitions are compared on the basis of the mutual coreference information. R is obtained by subtracting the conditioned entropy of GOLD given SYS (loss of information) from the entropy of GOLD. P is obtained by subtracting the conditioned entropy of SYS given GOLD (irrelevant information gains) from the SYS entropy. Both values are then normalized. This measure has been hardly used for reporting results of real systems, and it emerges from the results reported by Popescu-Belis (2000) that H is not superior to the other existing measures. Popescu-Belis concludes that each metric, by focusing on different aspects of the data, provides a different perspective on the quality of the system answer.

Rand index (Rand 1971). The Rand index is a general clustering evaluation metric that measures the similarity between two clusterings (i.e. partitions) by considering how each pair of data points is assigned in each clustering. Stated in coreference

[3] In the ACE evaluation program, mentions and entities in SYS that are not mapped onto any mention or entity in GOLD receive a false alarm penalty.

https://doi.org/10.1017/S135132491000029X Published online by Cambridge University Press

Table 3. *Performance of state-of-the-art coreference systems on ACE*

System	MUC-F	B^3-F	CEAF	ACE-value
		ACE-2		
All-singletons baseline	–	55.9	38.8	
One-entity baseline	76.5	17.3	21.7	
Luo *et al.* (2004)	80.7	77.0	73.2	89.8
Finkel and Manning (2008)	64.1	73.8		
Poon and Domingos (2008)	68.4	69.2	63.9	
Denis and Baldridge (2009)	70.1	72.7	66.2	
Ng (2009)	61.3		61.6	
		ACE-2004		
All-singletons baseline	–	59.0	41.8	
One-entity baseline	74.4	17.8	21.4	
Luo and Zitouni (2005)	86.0	83.7	82.0	91.6
Haghighi and Klein (2007)	63.3			
Bengtson and Roth (2008)	75.8	80.8	75.0	
Poon and Domingos (2008)	69.1	71.2	65.9	
Wick and McCallum (2009)	70.1	81.5		

terms, the Rand index equals the number of mention pairs that are either placed in an entity or assigned to separate entities in both GOLD and SYS, normalized by the total number of mention pairs in each partition. The motivations behind this measure are three (where we replace 'point' by 'mention', 'cluster' by 'entity', and 'clustering' by 'entity partition'): (1) every mention is unequivocally assigned to a specific entity; (2) entities are defined just as much by those points which they do not contain as by those mentions which they do contain; and (3) all mentions are of equal importance in the determination of the entity partition.

The only use of the Rand index for coreference resolution appears in Finkel and Manning (2008). 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 (Table 2). The extremely high number of mention pairs that are found in different entities in GOLD and SYS explains the high figures obtained by all systems reported in Finkel and Manning (2008), and by system response (d) in Table 1. Hence, the low discriminatory power of Rand. The BLANC measure that we introduce in Section 4 implements Rand in a way suited to the coreference problem.

It is often hard for researchers working on coreference resolution to make sense of the state of the art. Compare, for example, the scores shown in Table 3 that correspond to various systems[4] and two baselines: (1) all singletons (i.e. no coreference link is created, but each mention is considered to be a separate entity),

[4] Scores published here but missing in the original papers were computed by us from the authors' outputs.

https://doi.org/10.1017/S135132491000029X Published online by Cambridge University Press

and (2) one entity (i.e. all document mentions are clustered into one single entity). The only measure for which we have the results of all systems is MUC, but this is the one with the largest number of drawbacks, as evidenced by the one-entity baseline's high score. It is clear that the measures do not produce the same ranking of the systems, other than the fact that they all rank Luo *et al.* (2004) and Luo and Zitouni (2005) as the best systems for each data set. This sort of discrepancy makes it impossible in the long term to conduct research on this question: which measure should one trust, and why?

Apart from the pros and cons of each measure, the difficulty in comparing the performance of different coreference resolution systems is compounded by other factors, such as the use of true or system mentions and the use of different test sets (Stoyanov *et al.* 2009). Some systems in Table 3 are not directly comparable since testing on a different set of mentions or on a different data set is likely to affect scoring. Ng (2009) did not use true but system mentions, and Luo and Zitouni (2005) had access to the entire ACE-2004 formal test sets, while the remaining systems, due to licensing restrictions, were evaluated on only a portion of the ACE-2004 training set.

3.2 Desiderata for a coreference evaluation measure

Coreference is a type of clustering task, but it is special as each item in a cluster bears the same relationship, referential identity, with all other items in the same cluster, plus the fact that a large number of clusters are singletons. Thus, only two of the four formal constraints for clustering evaluation metrics pointed out by Amigó *et al.* (2009) apply to coreference. Amigó *et al.*'s (2009) formal constraints include: (1) cluster homogeneity, i.e. clusters should not mix items belonging to different categories; (2) cluster completeness, i.e. items belonging to the same category should be grouped in the same cluster; (3) rag bag, i.e. it is preferable to have clean clusters plus a cluster with miscellaneous items over having clusters with a dominant category plus additional noise; and (4) cluster size versus quantity, i.e. a small error in a large cluster is preferable to a large number of small errors in small clusters.

While the first two constraints undoubtedly hold for coreference resolution, the last two do not necessarily. What makes coreference resolution special with respect to other clustering tasks is the propagation of relations within an entity caused by the transitive property of coreference. That is to say, unlike regular clustering, where assigning a new item to a cluster is a mere question of classifying that item into a specific category, in coreference resolution assigning a new mention to an entity implies that the mention is coreferent with *all* other mentions that have been assigned to that same entity. Thus, the larger an entity is, the more coreferent links will be asserted for each new mention that is added.

To illustrate: To us, given the GOLD in Figure 4, the output produced by system S2 is not better than that produced by system S1, as it would follow from constraint (3). In fact, if the rag-bag entity contained more singletons, including an additional wrong singleton would make S2 even worse than S1. Similarly, in Figure 5, S2 is not better than S1, as suggested by constraint (4).

https://doi.org/10.1017/S135132491000029X Published online by Cambridge University Press

GOLD = { {Barack Obama, the president, Obama}, {Sarkozy}, {Berlin}, {the UN}, {today} }

S1 = { {Barack Obama, the president, Obama, Sarkozy}, {Berlin}, {the UN}, {today} }

S2 = { {Barack Obama, the president, Obama}, {Sarkozy, Berlin, the UN, today} }

Fig. 4. An example not satisfying constraint (3): The output S2 with a rag-bag cluster is equally preferable to S1.

GOLD = { {Barack Obama, the president, Obama}, {the French capital, Paris}, {the Democrats, the Democrats} }

S1 = { {Barack Obama, the president, Obama}, {the French capital}, {Paris}, {the Democrats}, {the Democrats} }

S2 = { {Barack Obama, the president}, {Obama}, {the French capital, Paris}, {the Democrats, the Democrats} }

Fig. 5. An example not satisfying constraint (4): The output S2 with a small error in a large cluster is equally preferable to S1.

Amigó *et al.* (2009) show that whereas B^3 satisfies all four constraints, measures based on counting pairs, such as the Rand index, satisfy only constraints (1) and (2). This is the reason why Rand is a good starting point for developing the BLANC measure for coreference resolution in Section 4. As described in Section 3.1, the three most important points that remain unsolved by the current coreference metrics are as follows:

(1) *Singletons.* Since including a mention in the wrong chain hurts P, a correct decision to NOT link a mention should be rewarded as well. Rewarding correctly identified singletons, however, needs to be moderate, leaving enough margin for the analysis of correctly identified multimention entities.
(2) *Boundary cases.* Special attention needs to be paid to the behavior of the evaluation measure when a system outputs (1) all singletons, or (2) one entity (i.e. all mentions are linked).
(3) *Number of mentions.* The longer the entity chain, the more coreferent mentions it contains, each mention inheriting the information predicated of the other mentions. Thus, a correct large entity should be rewarded more than a correct small entity, and a wrong large entity should be penalized more than a wrong small entity.

We suggest that a good coreference evaluation measure should conform to the following desiderata:

(1) Range from 0 for poor performance to 1 for perfect performance.
(2) Be monotonic: Solutions that are obviously better should obtain higher scores.

https://doi.org/10.1017/S135132491000029X Published online by Cambridge University Press

(3) Reward P more than R: Stating that two mentions are coreferent when they are not is more harmful than missing a correct coreference link.[5] Hence, the score should move closer to 1 as
   - More correct coreference links are found,
   - more correct singletons are found,
   - fewer wrong coreference links are made.

(4) Provide sufficiently fine scoring granularity to allow detailed discrimination between systems across the whole range [0, 1].

(5) As nearly as possible, maintain the same degree of scoring granularity throughout the whole range [0, 1].

4 BLANC: BiLateral assessment of noun-phrase coreference

In order to facilitate future research, we propose BLANC, a measure obtained by applying the Rand index (Rand 1971) to coreference and taking into account the above-mentioned problems and desiderata. The class-based methods 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 multimention entities. In addition, the MUC metric suffers from the essential problem that it does not explicitly reward correctly identified singletons, yet penalizes singletons when incorrectly included as part of a chain, while it is too lenient with penalizing wrong coreference links.

4.1 Implementing the Rand index for coreference evaluation

From what has been said in Section 3, 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 correctly handle singletons as well as to reward correct coreference chains commensurately with their length.[6] 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.

The Rand index (1) uses N_{00} (i.e. the number of mention pairs that are in the same entity in both GOLD and SYS) and N_{11} (i.e. the number of mention pairs that

[5] Although this is debatable, as it might depend on the application for which the coreference output is used, it is a widespread belief among researchers that P matters more than R in coreference resolution.

[6] We define a non-coreference link to hold between every two mentions that are deemed to NOT corefer.

https://doi.org/10.1017/S135132491000029X Published online by Cambridge University Press

Table 4. *The BLANC confusion matrix*

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

are in different entities in both GOLD and SYS) as agreement indicators between the two partitions GOLD and SYS. The value of Rand lies between 0 and 1, with 0 indicating that the two partitions do not agree on any pair of mentions and 1 indicating that the partitions are identical.

$$\text{Rand} = \frac{N_{00} + N_{11}}{N(N-1)/2} \tag{1}$$

BLANC borrows the 'bilateral' nature of Rand to take into consideration both coreference links (N_{00}) and non-coreference links (N_{11}), but modifies it such that every decision of coreferentiality is assigned equal importance. Thus, BLANC models coreference resolution better by addressing the significant imbalance between the number of coreferent mentions and singletons observed in the real data. Further, whereas class-based metrics need to address the fact 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 in the fact that the total number of links remains constant across GOLD and SYS.

4.1.1 Coreference and non-coreference links

BLANC is best explained considering the following two kinds of decisions:

(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 4 shows the 2 × 2 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.

https://doi.org/10.1017/S135132491000029X Published online by Cambridge University Press

Table 5. *The BLANC confusion matrix for the example in Figure 1*

		SYS		
		Coreference	Non-coreference	Sums
GOLD	Coreference	2	2	4
	Non-coreference	3	84	87
Sums		5	86	91

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

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

The total number of links in the SYS partition of d is the sum of the four possible types of links, and it equals L:

$$rc + wc + rn + wn = L$$

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 confusion matrix for the example in Figure 1 is shown in Table 5. As the text has fourteen mentions, the total number of links is ninety-one. The system correctly identifies two coreference links (m_5–m_{12}, m_7–m_9), and wrongly another three coreference links (m_4–m_6, m_7–m_{14}, m_9–m_{14}). Every right coreference link that is missed by the system necessarily produces a wrong non-coreference link (m_5–m_{14}, m_{12}–m_{14}). The rest are eighty-four right non-coreference links. The confusion matrix shows the balance between coreference and non-coreference links with respect to the gold partition.

The singleton problem pointed out in Section 3 becomes evident in Table 5: The number of non-coreference links is much higher than the number of coreference links. The class imbalance problem of coreference resolution causes that if the Rand index is applied as originally defined by Rand (1971), it concentrates in a small interval near 1 with hardly any discriminatory power. A chance-corrected Rand index has been proposed (Hubert and Arabie 1985), but it is of no use for the coreference problem, given that the computation of expectation only depends on the number of pairs in the same cluster, thus ignoring singletons.

In order to take into account the under-representation of coreference links in the final BLANC score, we compute P, R, and F separately for the two types of links (coreference and non-coreference) and then average them for the final score. The definition of BLANC is shown in Table 6. In BLANC, both coreference and non-coreference links contribute to the final score, but neither more than 50 per cent. 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

https://doi.org/10.1017/S135132491000029X Published online by Cambridge University Press

Table 6. *Definition: formula for BLANC*

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 7. *Performance of example in Figure 1*

MUC-F	B^3-F	CEAF	BLANC
57.14	86.76	85.71	70.78

F-scores. Applying the Rand index, the novelty of BLANC resides in putting equal emphasis on coreference and non-coreference links. Table 7 shows the different measures under discussion for the example in Figure 1.

4.1.2 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 6.

- 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

https://doi.org/10.1017/S135132491000029X Published online by Cambridge University Press

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, can introduce a parameter that enables the user to change the relative weights given to coreference and non-coreference links. We provide details in the following section.

4.1.3 The α parameter

After analyzing several coreferentially annotated corpora, we found that the average text contains between 60 per cent and 80 per cent singletons (depending on the coding scheme). Thus, simply averaging the coreference and non-coreference scores seems to be the best decision. However, given extraordinary cases like the one presented at the end of Section 4.1.2 or for those researchers that consider it to be convenient, we present the weighted version of BLANC:

$$\text{BLANC}_\alpha = \alpha \text{F}_c + (1 - \alpha)\text{F}_n$$

BLANC_α lets users choose the weights they want to put on coreference and non-coreference links. In the default version of BLANC (Table 6), $\alpha = 0.5$. Setting α closer to 1 will give a larger weight to coreference links, while setting α closer to 0 will have the opposite effect. For the problematic near-boundary case in which all links but one are non-coreferent in GOLD, evaluating with $\text{BLANC}_{\alpha=0.1}$ will be much less severe than evaluating with the default BLANC.

4.2 Identification of mentions

An additional drawback that has been pointed out for the class-based metrics like B^3 and CEAF is their assumption of working with true mentions, ignoring the problem of evaluating end-to-end systems, where some mentions in SYS might not be correct; i.e. might not be mapped onto any mention in GOLD and *vice versa*. These are called 'twinless' mentions by Stoyanov *et al.* (2009). Bengtson and Roth (2008) simply discard twinless mentions, and Rahman and Ng (2009) limit to removing only those twinless system mentions that are singletons, as in these cases no penalty should be applied. Recently, Cai and Strube (2010) have proposed two variants of B^3 and CEAF that put twinless gold mentions into SYS as singletons and discard singleton twinless system mentions. In order to calculate P, wrongly resolved twinless system mentions are put into GOLD; to calculate R, only the gold entities are considered.

We agree that proper evaluation of a coreference system should take into account true versus system mentions. However, the mention identification task strictly belongs to syntax as it is closely related to the problem of identifying noun-phrase boundaries, followed by a filtering step in which only referential noun phrases are retained. It is clearly distinct from coreference resolution, whose goal is to link those noun phrases that refer to the same entity. One single metric giving the overall result for the two tasks together is obscure in that it is not informative as to whether a system is very good at identifying coreference links but poor at identifying mention boundaries, or *vice versa*. Therefore, instead of merging the two tasks, we propose to consider

https://doi.org/10.1017/S135132491000029X Published online by Cambridge University Press

mention identification as its own task and separate its evaluation from that of coreference resolution (Popescu-Belis *et al.* 2004). In brief, a measure for each problem is as follows:

- *Mention identification.* This evaluation computes the correctness of the mentions that are being resolved, regardless of the structure of coreference links. Standard P and R are computed to compare the sets of mentions of GOLD and SYS. P is defined as the number of common mentions between GOLD and SYS divided by the number of system mentions; R is defined as the number of common mentions between GOLD and SYS divided by the number of true mentions. Two versions for the matching module are possible:
  - — Strict matching: A system mention is considered to be correctly identified when it exactly matches the corresponding gold mention.
  - — Lenient matching: A system mention is considered to be correctly identified when it matches at least the head of the corresponding gold mention (and does not include any tokens outside the gold mention).[7]
- *Correctness of coreference.* This evaluation computes the correctness of the coreference links predicted between the mentions shared by GOLD and SYS. The BLANC measure is applied to this set of correctly recognized mentions.

In this way, it might be possible to improve under-performing systems by combining, for instance, the strengths of a system that obtains a high coreference score but a low mention-identification score with the strengths of a system that performs badly in coreference resolution but successfully in the identification of mentions. Similarly, one should not be led to believe that improving the set of coreference features will necessarily result in higher scores, as the system's mention-identification score might reveal that the underlying problem is a poor detection of true mentions.

5 Discriminative power

This section empirically demonstrates the power of BLANC by comparing its scores with those of MUC, B^3, CEAF, and the Rand index on both artificial and real gold/system partitions. The insight provided by BLANC is free of the problems noted in Section 3. This being said, we need to draw attention to the difficulty of agreeing on what 'correctness' means in coreference resolution. People's intuitions about the extreme boundary cases largely coincide, but those about intermediate cases, which are harder to evaluate, might differ considerably due to the complex trade-off between P and R. Thus, the discussion that follows is based on what we believe to be the best ranking of system responses according to our intuitions and to our experience in coreference annotation and resolution.

[7] Lenient matching is equivalent to the MIN attribute used in the MUC guidelines (Hirschman and Chinchor 1997) to indicate the minimum string that the system under evaluation must include.

https://doi.org/10.1017/S135132491000029X Published online by Cambridge University Press

Table 8. *Different system responses for a gold standard* $Gold_1$

Response	Output
$Gold_1$	(1) (2) (3) (4) (5) (6) (7) (8) (9) (10) (11) (12) (13) (14) (15) (16) (17) (18) (19) (20) (21) (22) (23) (24) (25) (26) (27) (28) (29) (30) (31) (32) (33) (34) (35) (36) (37) (38) (39) (40) (41) (42) (43) (44) (45) (46) (47) (48) (49) (50) (51) (52) (53) (54) (55) (56) (57) (58) (59) (60) (61) **(62, 63, 64, 65) (66, 67, 68) (69, 70)**
System A	**(1, 2)** (3) (4) (5) (6) (7) (8) (9) (10) (11) (12) (13) (14) (15) (16) (17) (18) (19) (20) (21) (22) (23) (24) (25) (26) (27) (28) (29) (30) (31) (32) (33) (34) (35) (36) (37) (38) (39) (40) (41) (42) (43) (44) (45) (46) (47) (48) (49) (50) (51) (52) (53) (54) (55) (56) (57) (58) (59) (60) (61) **(62, 63, 64, 65) (66, 67, 68) (69, 70)**
System B	**(1, 62, 63, 64, 65)** (2) (3) (4) (5) (6) (7) (8) (9) (10) (11) (12) (13) (14) (15) (16) (17) (18) (19) (20) (21) (22) (23) (24) (25) (26) (27) (28) (29) (30) (31) (32) (33) (34) (35) (36) (37) (38) (39) (40) (41) (42) (43) (44) (45) (46) (47) (48) (49) (50) (51) (52) (53) (54) (55) (56) (57) (58) (59) (60) (61) **(66, 67, 68) (69, 70)**
System C	(1) (2) (3) (4) (5) (6) (7) (8) (9) (10) (11) (12) (13) (14) (15) (16) (17) (18) (19) (20) (21) (22) (23) (24) (25) (26) (27) (28) (29) (30) (31) (32) (33) (34) (35) (36) (37) (38) (39) (40) (41) (42) (43) (44) (45) (46) (47) (48) (49) (50) (51) (52) (53) (54) (55) (56) (57) (58) (59) (60) (61) **(62, 63, 64, 65)** (66) (67) (68) **(69, 70)**
System D	(1) (2) (3) (4) (5) (6) (7) (8) (9) (10) (11) (12) (13) (14) (15) (16) (17) (18) (19) (20) (21) (22) (23) (24) (25) (26) (27) (28) (29) (30) (31) (32) (33) (34) (35) (36) (37) (38) (39) (40) (41) (42) (43) (44) (45) (46) (47) (48) (49) (50) (51) (52) (53) (54) (55) (56) (57) (58) (59) (60) (61) **(62,63,64,65,66,67,68) (69,70)**
System E	**(1,62,63)** (2) (3) (4) (5) (6) (7) (8) (9) (10) (11) (12) (13) (14) (15) (16) (17) (18) (19) (20) (21) (22) (23) (24) (25) (26) (27) **(28,64,65)** (29) (30) (31) (32) (33) (34) (35) (36) (37) (38) (39) (40) (41) (42) (43) (44) (45) (46) (47) (48) (49) (50) (51) (52) (53) (54) (55) (56) (57) (58) (59) (60) (61) **(66,67,68) (69,70)**
System F	**(1,62)** (2) (3) **(4,63)** (5) (6) (7) (8) (9) (10) (11) (12) (13) (14) (15) (16) (17) (18) (19) (20) (21) (22) (23) (24) (25) (26) (27) **(28,64)** (29) (30) (31) (32) (33) (34) (35) (36) (37) (38) (39) (40) (41) (42) (43) (44) (45) (46) (47) (48) (49) (50) (51) (52) (53) (54) (55) (56) **(57,65)** (58) (59) (60) (61) **(66,67,68) (69,70)**
System G	All singletons
System H	One entity

5.1 Results on artificial data

We take the gold partition in the first row of Table 8 as a working example. It is representative of a real case: it contains seventy mentions, 95 per cent singleton entities, a two-mention entity, a three-mention entity, and a four-mention entity. Each

https://doi.org/10.1017/S135132491000029X Published online by Cambridge University Press

Table 9. *Decomposition of the system responses in Table 8*

System	#entities	#singletons	*rc*	*rn*	*wc*	*wn*
A	63	59	10	2,404	1	0
B	63	60	10	2,401	4	0
C	66	64	7	2,405	0	3
D	63	61	10	2,393	12	0
E	63	59	6	2,401	4	4
F	63	57	4	2,401	4	6
G	70	70	0	2,405	0	10
H	1	0	10	0	2,405	0

number represents a different mention; parentheses identify entities (i.e. they group mentions that corefer); and multimention entities are highlighted in bold. Table 8 also contains eight sample responses – output by different hypothetical coreference resolution systems – that contain different types of errors. See the decomposition into BLANC's four types of links in Table 9, a quantitative representation of the quality of the systems given in Table 8. The responses are ranked in order of quality, from the most accurate response to the least (response A is better than response B, B is better than C, and so on, according to our intuitions[8]).

System A commits only one P error by linking two non-coreferent mentions; system B looks similar to A but is worse in that a singleton is clustered in a four-mention entity, thus producing not one but four P errors. System C exhibits no P errors but is weak in terms of R, as it fails to identify a three-mention entity. Although system D is clean in terms of R, it suffers from a severe P problem due to the fusion of the three- and four-mention entities in one large entity. System E is worse than the previous responses, as it shows both P and R errors: the four-mention entity is split into two and a singleton is added to both of them. System F worsens the previous output by completely failing to identify the four-mention entity and creating four incorrect two-mention entities. Finally, systems G and H represent the two boundary cases, the former being preferable to the latter, since at least it gets the large number of singletons, while the latter has a serious problem in P.

The performance of these system responses according to different measures is given in Tables 10 and 11. In them, we can see how BLANC addresses the three problems noted in Section 3.2.

(1) *Singletons.* The BLANC score decreases as the response quality decreases. It successfully captures the desired ranking, so does CEAF (although with fewer distinctions, see the 'number of mentions' problem below), and so does B^3 if we leave aside the boundary responses G and H. BLANC, however, shows a much wider interval (from 97.61 to 49.90 per cent) than CEAF (from 98.57 to

[8] Readers and reviewers of this section frequently comment that this ranking is not clearly apparent; other variations seem equally good. We concede this readily. We argue that in cases when several rankings seem intuitively equivalent to people, one can accept the ranking of a metric, as long as it assigns relatively close scores to the equivalent cases.

https://doi.org/10.1017/S135132491000029X Published online by Cambridge University Press

Table 10. *Performance of the systems in Table 8*

System	MUC-F	B^3-F	CEAF	Rand	BLANC
A	92.31	99.28	98.57	99.96	97.61
B	92.31	98.84	98.57	99.83	91.63
C	80.00	98.55	97.14	99.88	91.15
D	92.31	97.49	95.71	99.50	81.12
E	76.92	96.66	95.71	99.67	79.92
F	46.15	94.99	94.29	99.59	72.12
G	–	95.52	91.43	99.59	49.90
H	16.00	3.61	5.71	0.41	0.41

Table 11. *P and R scores for the systems in Table 8*

	MUC		B^3		CEAF	BLANC	
System	P	R	P	R	P/R	P	R
A	85.71	100.00	98.57	100.00	98.57	95.45	99.98
B	85.71	100.00	97.71	100.00	98.57	85.71	99.92
C	100.00	66.67	100.00	97.14	97.14	99.94	85.00
D	85.71	100.00	95.10	100.00	95.71	72.73	99.75
E	71.43	83.33	96.19	97.14	95.71	79.92	79.92
F	42.86	50.00	94.29	95.71	94.29	74.88	69.92
G	–	–	100.00	91.43	91.43	49.79	50.00
H	8.70	100.00	1.84	100.00	5.71	0.21	50.00

91.43 per cent) and B^3 (from 99.28 to 94.99 per cent), thus providing a larger margin of variation, and a finer granularity. The singleton problem is solved by rewarding the total number of correct singletons as much as the total number of correct mentions in multimention entities. Note that the original Rand index makes it impossible to discriminate between systems and it does not even rank them as intuitively expected.

(2) *Boundary cases.* MUC fails to capture the fact that the all-singletons response G is better than the one-entity response H. On the other hand, B^3 and CEAF give a score close to 0 per cent for H, yet close to 100 per cent for G. It is counterintuitive that a *coreference* resolution system that outputs as many entities as mentions – meaning that it is doing nothing – gets such a high score. BLANC successfully handles the boundary responses by setting an upper bound of 50 percent on R.

(3) *Number of mentions.* The fact that MUC and CEAF give the same score to responses A and B shows their failure at distinguishing that the latter is more harmful than the former, as it creates more false coreference links. Namely, the information predicated of mention 1 is extended to mentions 61, 62, 63, and 64, and reciprocally mention 1 gets all the information predicated of mentions 61, 62, 63, and 64. Similarly, CEAF does not distinguish response D from E.

https://doi.org/10.1017/S135132491000029X Published online by Cambridge University Press

Table 12. *Different system responses for a gold standard* $Gold_2$

Response	Output
$Gold_2$	(1) (2) (3) (4) (5) (6) (7) (8) (9) (10) (11) (12) (13) (14) (15) (16) **(17, 18)**
System A	(1) (2) (3) (4) (5) (6) (7) (8) (9) (10) (11) (12) (13) (14) (15) (16) (17) (18)
System B	(1) (2) (3) (4) (5) (6) (7) (8) (9) (10) (11) (12) (13) (14) (15) **(16, 17)** (18)
System C	(1) (2) (3) (4) (5) (6) (7) (8) (9) (10) (11) (12) (13) (14) **(15, 16)** (17) (18)
System D	(1) (2) (3) (4) (5) (6) (7) (8) (9) (10) (11) (12) (13) (14) **(15, 16) (17, 18)**

Table 13. *Performance for the systems in Table 12*

System	MUC-F	B^3-F	CEAF	$BLANC_{\alpha=0.5}$	$BLANC_{\alpha=0.2}$	$BLANC_{\alpha=0.1}$
A	–	97.14	94.44	49.84	79.74	89.70
B	0.00	94.44	94.44	49.67	79.47	89.41
C	0.00	94.44	88.89	49.67	79.47	89.41
D	66.67	97.14	94.44	83.17	93.07	96.37

In contrast, BLANC can discriminate between these responses because its reward of multimention entities is correlated with the number of coreference links contained in them.

The constructed example in Table 12 serves to illustrate BLANC's major weakness, which we have discussed at the end of Section 4.1.2. Performance is presented in Table 13. Notice the enormous leap between the $BLANC_{\alpha=0.5}$ score for system D and the other three. This is due to the fact that partitions A, B, and C contain no right coreference link, and so BLANC is equal to the correctness of non-coreference links divided by two. The α parameter introduced in Section 4.1.3 is especially adequate for this type of cases. The difference in the scores for D and the rest of systems diminishes when $\alpha = 0.2$ or $\alpha = 0.1$ (the two last columns).

This same example, in fact, reveals weaknesses of all the measures. Owing to the fact that the MUC score does not reward correctly identified singletons, it is not able to score the first three responses, thus showing even a larger rise in response D. The B^3 and CEAF measures score responses A and D the same, but only the latter succeeds in identifying the only coreference link that exists in the truth – a very relevant fact given that the ultimate goal of a coreference resolution system is not outputting only singletons (as system A does) but solving coreference. Finally, it is puzzling that CEAF considers response B to be appreciably better than response C – they are scored the same by B^3 and BLANC. This is a weakness due to CEAF's one-to-one alignment: In B, the three final entities find a counterpart in the gold standard, whereas in C, only one of the two final entities gets mapped.

5.2 Results on real data

In order not to reach conclusions solely derived from constructed toy examples, we run a prototype learning-based coreference resolution system – inspired by Soon

https://doi.org/10.1017/S135132491000029X Published online by Cambridge University Press

Table 14. *Different coreference resolution models run on ACE-2004*

Resolution model	MUC-F	B^3-F	CEAF	BLANC
A. All-singletons baseline	–	67.51	50.96	48.61
B. Head-match baseline	52.93	76.60	66.46	66.35
C. Strong match	64.69	75.56	**70.63**	**73.76**
D. Best match	61.60	**76.76**	69.19	71.98
E. Weak match	**70.34**	70.24	64.00	66.50

Table 15. *Decomposition of the system responses in Table 14*

Resolution model	#entities	#singletons	*rc*	*rn*	*wc*	*wn*
A. All-singletons baseline	1,464	1,464	0	39,672	0	2,272
B. Head-match baseline	1,124	921	506	39,560	112	1,766
C. Strong match	735	400	1,058	38,783	889	1,214
D. Best match	867	577	870	39,069	603	1,402
E. Weak match	550	347	1,757	34,919	4,753	515

et al. (2001), Ng and Cardie (2002), and Luo *et al.* (2004) – on 33 documents of the ACE-2004 corpus. A total of five different resolution models are tried to enable a richer analysis and comparison between the different evaluation metrics. The results are presented in Table 14. For a detailed analysis we address the reader to Recasens and Hovy (2010).

The first two are baselines that involve no learning: model A is the all-singletons baseline, and B clusters in the same entity all the mentions that share the same head. In C, D, and E, a pairwise coreference classifier is learnt (i.e. given two mentions, it classifies them as either coreferent or non-coreferent). In C and D, whenever the classifier considers two mentions to be coreferent and one of them has already been clustered in a multimention entity, the new mention is only clustered in that same entity if all pairwise classifications with the other mentions of the entity are also classified as coreferent. The difference between C and D lies in the initial mention pairs that form the basis for the subsequent process: C takes the first mention in textual order that is classified as coreferent with the mention under consideration, while D takes the mention that shows the highest confidence among the previous. E is a simplified version of C that performs no additional pairwise checks.

The best way to judge the quality of each response is to look at the actual data, but space limitations make this impossible. However, we can gain an approximation by looking at Table 15, which shows the number of entities output by each system and how many are singletons, as well as the number of correct and incorrect links of each type. Note that high numbers in the *wc* column indicate poor P, whereas high numbers in the *wn* column indicate poor R. Although the trade-off between P and R makes it hard to reach a conclusion as to whether C or D should be ranked first, the low quality of A, and especially E, is an easier conclusion to reach. The head-match baseline achieves high P but low R.

https://doi.org/10.1017/S135132491000029X Published online by Cambridge University Press

Table 16. *Performance of state-of-the-art systems on ACE according to BLANC*

System	MUC	B^3	CEAF	ACE-value	BLANC
			ACE-2		
All-singletons baseline	–	55.9	38.8		**47.8**
One-entity baseline	76.5	17.3	21.7		**7.8**
Luo *et al.* (2004)	80.7	77.0	73.2	89.8	**77.2**
			ACE-2004		
All-singletons baseline	–	59.0	41.8		**48.1**
One-entity baseline	74.4	17.8	21.4		**7.0**
Luo and Zitouni (2005)	86.0	83.7	82.0	91.6	**81.4**
Bengtson and Roth (2008)	75.8	80.8	75.0		**75.6**

If we go back to Table 14, we can see that no two measures produce the same ranking of systems. The severe problems behind the MUC score are again manifested: it ranks model E first because it identifies a high number of coreference links, despite containing many incorrect ones. This model produces an output that is not satisfactory because it tends to overmerge. The fact that B^3 ranks D and B first indicates its focus on P rather than R. Thus, B^3 tends to score best those models that are more conservative and that output a large number of singletons. Finally, CEAF and BLANC agree in ranking C the best. An analysis of the data also supports the idea that strong match achieves the best trade-off between P and R.

Similar problems with the currently used evaluation metrics were also shown by the six systems that participated in the SemEval-2010 Task 1 on 'Coreference Resolution in Multiple Languages' (Recasens *et al.* 2010), where the BLANC measure was publicly used for the first time. Unlike ACE, mentions were not restricted to any semantic type, and the B^3 and CEAF scores for the all-singletons baseline were hard to beat even by the highest performing systems. The BLANC scores, in contrast, tended to stay low regardless of the number of singletons in the corpus. However, it was not possible to draw definite conclusions about the SemEval shared task because each measure ranked the participating systems in a different order.

Finally, in Table 16 we reproduce Table 3 adding the BLANC score for the performance of state-of-the-art systems and the all-singletons and one-entity baselines. We can only include the results for those systems whose output responses were provided to us by the authors. It is worth noting that BLANC is closer to B^3 when using the ACE-2 corpus but closer to CEAF when using the ACE-2004 corpus, which is probably due to the different distribution of singletons and multimention entities in each corpus. Knowing the state of the art in terms of BLANC will enable future researchers on coreference resolution to compare their performance against these results.

5.3 Plots

A graph plotting the BLANC slope as the percentage of correct coreference links (*rc*) increase is depicted in Figure 6, where the slopes of B^3, CEAF, and MUC are also

https://doi.org/10.1017/S135132491000029X Published online by Cambridge University Press

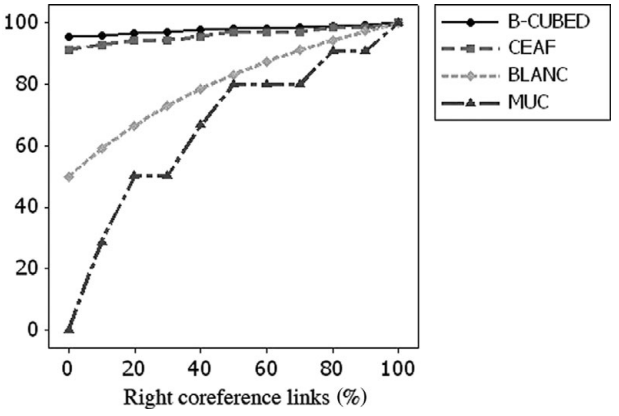


Fig. 6. The BLANC score curve as the number of right coreference links increases.

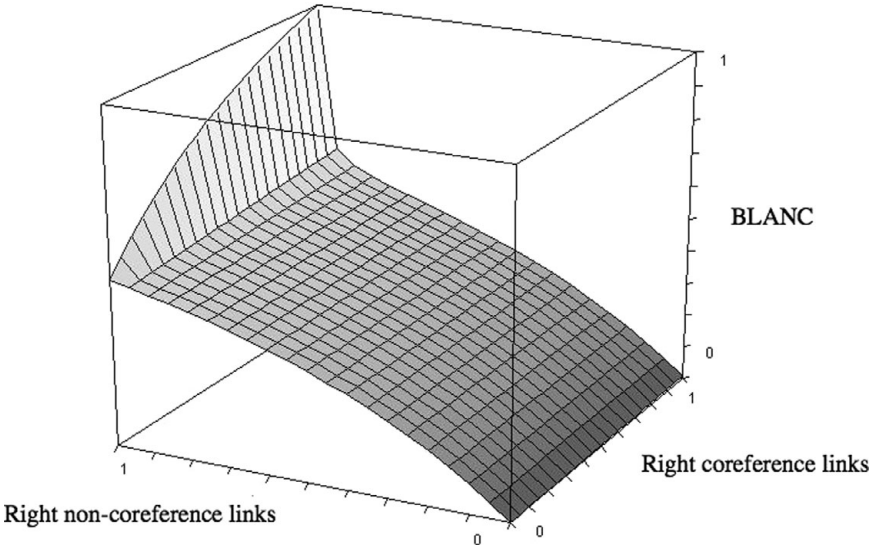


Fig. 7. The BLANC score surface as a function of right coreference and right non-coreference links, for data from Table 8.

plotted. The curve slope for BLANC gradually increases, and stays between the other measures, higher than MUC but lower than B^3 and CEAF, which show an almost flat straight line. The 'pinching' of scores close to 100 per cent by B^3 and CEAF is clearly apparent. A coreference resolution system can obtain very high B^3 and CEAF scores (due to the high number of singletons that are present in the gold partition), leaving a too small margin for the evaluation of coreference proper.

We illustrate in Figure 7 the dependency of the BLANC score on degrees of coreference and non-coreference. Figure 7 plots the scores for the example in Table 8. The left rear face of the cube – where the right non-coreference (i.e. *rn*) level is a constant 1 and right coreference (*rc*) ranges from zero to 1 – displays the BLANC curve from Figure 6. The front face of the cube shows how – for a

https://doi.org/10.1017/S135132491000029X Published online by Cambridge University Press

constant right coreference of 1 – the BLANC score ranges from near zero to 0.5 as right non-coreference ranges from zero to 1. The bend in the surface occurs due to the asymmetry in the number of true coreferences: the smaller the proportion of coreference links to non-coreference links, the sharper the bend and the closer it is to the left face. Systems must achieve correctness of almost all coreference *and* non-coreference links to approach the steep curve.

6 Conclusion

This paper seeks to shed light on the problem of coreference resolution evaluation by providing desiderata for coreference evaluation measures, pointing out the strong and weak points of the main measures that have been used, and proposing the BLANC metric, an implementation of the Rand index for coreference, to provide some further insight on a system's performance. The decomposition into four types of links gives an informative analysis of a system. BLANC fulfills the five desiderata and addresses to some degree the reported shortcomings of the existing measures. Despite its shortcomings, discussed in Sections 4.1.2 and 5.1, it overcomes the problem of singletons, which we illustrate here for the first time.

The simplicity of the BLANC measure derives from the fact that the sum of the coreference and non-coreference links in the gold and system partitions is the same. Unlike the Rand index, BLANC is the average of two F-scores, one for the coreference links and the other for the non-coreference links. Being two harmonic means, each F-score is lower than the normal average of P- and R-score – unless both are high. As a result, a coreference resolution system has to get *both* P and R for both coreference and non-coreference correct simultaneously to score well under BLANC. Although coreference and non-coreference are duals, ignoring one of the two halves means that some portion of the full link set remains unconsidered by the existing measures.

Tests on artificial and real data show that no evaluation measure is free of weaknesses and so at least two scoring measures should be used when evaluating a system. We argue that BLANC is consistent and achieves a good compromise between P and R. Its discriminative power – higher with respect to currently used metrics like MUC and B^3 – facilitates comparisons between coreference resolution systems.

Finally, this paper illustrates the need for a fuller comparison of all the evaluation measures, considering corrections required for chance variation, typical variances of scores under different conditions and data sizes, etc. Such a study has not yet been done for any of the measures, and could make a major contribution to the growing understanding of evaluation in the various branches of natural language engineering in general.

Acknowledgments

We would like to thank the anonymous reviewers for their helpful questions and comments. We are also indebted to Aron Culotta, Hal Daumé III, Jenny Finkel,

https://doi.org/10.1017/S135132491000029X Published online by Cambridge University Press

Aria Haghighi, Xiaoqiang Luo, Andrew McCallum, Vincent Ng, Hoifung Poon, and Nicholas Rizzolo for answering our request to recompute the performance of their coreference resolution systems with other metrics and/or providing us their system responses. Many thanks to Edgar Gonzàlez for implementation solutions.

This research was partially supported by the Spanish Ministry of Education through an FPU scholarship (AP2006-00994) and the TEXT-Knowledge 2.0 Project (TIN2009-13391-C04-04).

References

Amigó, E., Gonzalo, J., Artiles, J., and Verdejo, F. 2009. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information Retrieval* **12**(4): 461–486.

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

Bengtson, E., and Roth, D. 2008. Understanding the value of features for coreference resolution. In *Proceedings of EMNLP*, Honolulu, HI, pp. 294–303.

Cai, J., and Strube, M. 2010. Evaluation metrics for end-to-end coreference resolution systems. In *Proceedings of SIGDIAL*, University of Tokyo, Japan, pp. 28–36.

Choi, Y., and Cardie, C. 2007. Structured local training and biased potential functions for conditional random fields with application to coreference resolution. In *Proceedings of HLT-NAACL*, Rochester, NY, pp. 65–72.

Culotta, A., Wick, M., Hall, R., and McCallum, A. 2007. First-order probabilistic models for coreference resolution. In *Proceedings of HLT-NAACL*, Rochester, NY, pp. 81–88.

Daumé III, H., and Marcu, D. 2005. A large-scale exploration of effective global features for a joint entity detection and tracking model. In *Proceedings of HLT-EMNLP*, Vancouver, Canada, pp. 97–104.

Denis, P., and Baldridge, J. 2009. Global joint models for coreference resolution and named entity classification. *Procesamiento del Lenguaje Natural* **42**: 87–96.

Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., and Weischedel, R. 2004. The automatic content extraction (ACE) program – tasks, data and evaluation. In *Proceedings of LREC*, Lisbon, Portugal, pp. 837–840.

Finkel, J. R., and Manning, C. D. 2008. Enforcing transitivity in coreference resolution. In *Proceedings of ACL-HLT*, Columbus, OH, pp. 45–48.

Haghighi, A., and Klein, D. 2007. Unsupervised coreference resolution in a nonparametric Bayesian model. In *Proceedings of ACL*, Prague, Czech Republic, pp. 848–855.

Haghighi, A., and Klein, D. 2009. Simple coreference resolution with rich syntactic and semantic features. In *Proceedings of EMNLP*, Suntec, Singapore, pp. 1152–1161.

Hirschman, L., and Chinchor, N. 1997. MUC-7 coreference task definition – version 3.0. In *Proceedings of MUC-7*. Washington, DC.

Hubert, L., and Arabie, P. 1985. Comparing partitions. *Journal of Classification* **2**(1): 193–218.

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

Luo, X., Ittycheriah, A., Jing, H., Kambhatla, N., and Roukos, S. 2004. A mention-synchronous coreference resolution algorithm based on the Bell tree. In *Proceedings of ACL*, Barcelona, Spain, pp. 21–26.

Luo, X., and Zitouni, I. 2005. Multi-lingual coreference resolution with syntactic features. In *Proceedings of HLT-EMNLP*, Vancouver, Canada, pp. 660–667.

Ng, V. 2009. Graph-cut-based anaphoricity determination for coreference resolution. In *Proceedings of NAACL-HLT*, Boulder, CO, pp. 575–583.

Ng, V., and Cardie, C. 2002. Improving machine learning approaches to coreference resolution. In *Proceedings of ACL*, Philadelphia, PA, pp. 104–111.

https://doi.org/10.1017/S135132491000029X Published online by Cambridge University Press

Poon, H., and Domingos, P. 2008. Joint unsupervised coreference resolution with Markov logic. In *Proceedings of EMNLP*, Honolulu, HI, pp. 650–659.

Popescu-Belis, A. 2000. Évaluation numérique de la résolution de la référence: critiques et propositions. *T.A.L.: Traitement automatique de la langue* **40**(2): 117–146.

Popescu-Belis, A., Rigouste, L., Salmon-Alt, S., and Romary, L. 2004. Online evaluation of coreference resolution. In *Proceedings of LREC*, Lisbon, Portugal, pp. 1507–1510.

Rahman, A., and Ng, V. 2009. Supervised models for coreference resolution. In *Proceedings of EMNLP*, Suntec, Singapore, pp. 968–977.

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

Recasens, M., and Hovy, E. 2010. Coreference resolution across corpora: languages, coding schemes, and preprocessing information. In *Proceedings of ACL*, Uppsala, Sweden, pp. 1423–1432.

Recasens, M., Màrquez, L., Sapena, E., Martí, M. A., Taulé, M., Hoste, V., Poesio, M., and Versley, Y. 2010. SemEval-2010 task 1: coreference resolution in multiple languages. In *Proceedings of the Fifth International Workshop on Semantic Evaluation (SemEval 2010)*, Uppsala, Sweden, pp. 1–8.

Recasens, M., and Martí, M. A. 2010. AnCora-CO: coreferentially annotated corpora for Spanish and Catalan. *Language Resources and Evaluation* **44**(4): 315–345.

Soon, W. M., Ng, H. T., and Lim, D. C. Y. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics* **27**(4): 521–544.

Stoyanov, V., Gilbert, N., Cardie, C., and Riloff, E. 2009. Conundrums in noun phrase coreference resolution: making sense of the state-of-the-art. In *Proceedings of ACL-IJCNLP*, Suntec, Singapore, pp. 656–664.

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

Wick, M., and McCallum, A. 2009. Advances in learning and inference for partition-wise models of coreference resolution. Technical Report UM-CS-2009-028, Department of Computer Science, University of Massachusetts.

Yang, X., Su, J., Lang, J., Tan, C. L., Liu, T., and Li, S. 2008. An entity-mention model for coreference resolution with inductive logic programming. In *Proceedings of ACL-HLT*, Columbus, OH, pp. 843–851.

https://doi.org/10.1017/S135132491000029X Published online by Cambridge University Press