An Extension of BLANC to System Mentions

Xiaoqiang Luo
Google Inc.
111 8th Ave, New York, NY 10011
xql@google.com

Sameer Pradhan
Harvard Medical School
300 Longwood Ave., Boston, MA 02115
sameer.pradhan@childrens.harvard.edu

Marta Recasens
Google Inc.
1600 Amphitheatre Pkwy,
Mountain View, CA 94043
recasens@google.com

Eduard Hovy
Carnegie Mellon University
5000 Forbes Ave.
Pittsburgh, PA 15213
hovy@cmu.edu

Abstract

BLANC is a link-based coreference evaluation metric for measuring the quality of coreference systems on gold mentions. This paper extends the original BLANC ("BLANC-gold" henceforth) to system mentions, removing the gold mention assumption. The proposed BLANC falls back seamlessly to the original one if system mentions are identical to gold mentions, and it is shown to strongly correlate with existing metrics on the 2011 and 2012 CoNLL data.

1 Introduction

Coreference resolution aims at identifying natural language expressions (or mentions) that refer to the same entity. It entails partitioning (often imperfect) mentions into equivalence classes. A critically important problem is how to measure the quality of a coreference resolution system. Many evaluation metrics have been proposed in the past two decades, including the MUC measure (Vilain et al., 1995), B-cubed (Bagga and Baldwin, 1998), CEAF (Luo, 2005) and, more recently, BLANC-gold (Recasens and Hovy, 2011). B-cubed and CEAF treat entities as sets of mentions and measure the agreement between key (or gold standard) entities and response (or system-generated) entities, while MUC and BLANC-gold are link-based.

In particular, MUC measures the degree of agreement between key coreference links (i.e., links among mentions within entities) and response coreference links, while non-coreference links (i.e., links formed by mentions from different entities) are not explicitly taken into account. This leads to a phenomenon where coreference systems outputting large entities are scored more favorably than those outputting small entities (Luo, 2005). BLANC (Recasens and Hovy, 2011), on the other hand, considers both coreference links and non-coreference links. It calculates recall, precision and F-measure separately on coreference and non-coreference links in the usual way, and defines the overall recall, precision and F-measure as the mean of the respective measures for coreference and non-coreference links.

The BLANC-gold metric was developed with the assumption that response mentions and key mentions are identical. In reality, however, mentions need to be detected from natural language text and the result is, more often than not, imperfect: some key mentions may be missing in the response, and some response mentions may be spurious—so-called "twinless" mentions by Stoyanov et al. (2009). Therefore, the identical-mention-set assumption limits BLANC-gold's applicability when gold mentions are not available, or when one wants to have a single score measuring both the quality of mention detection and coreference resolution. The goal of this paper is to extend the BLANC-gold metric to imperfect response mentions.

We first briefly review the original definition of BLANC, and rewrite its definition using set notation. We then argue that the gold-mention assumption in Recasens and Hovy (2011) can be lifted without changing the original definition. In fact, the proposed BLANC metric subsumes the original one in that its value is identical to the original one when response mentions are identical to key mentions.

The rest of the paper is organized as follows. We introduce the notions used in this paper in Section 2. We then present the original BLANC-gold in Section 3 using the set notation defined in Section 2. This paves the way to generalize it to

Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 24–29,
Baltimore, Maryland, USA, June 23-25 2014.

imperfect system mentions, which is presented in Section 4. The proposed BLANC is applied to the CoNLL 2011 and 2012 shared task participants, and the scores and its correlations with existing metrics are shown in Section 5.

2 Notations

To facilitate the presentation, we define the notations used in the paper.

We use *key* to refer to gold standard mentions or entities, and *response* to refer to system mentions or entities. The collection of *key* entities is denoted by $K = \{k_i\}_{i=1}^{|K|}$, where k_i is the i^{th} key entity; accordingly, $R = \{r_j\}_{j=1}^{|R|}$ is the set of *response* entities, and r_j is the j^{th} response entity. We assume that mentions in $\{k_i\}$ and $\{r_j\}$ are unique; in other words, there is no duplicate mention.

Let $C_k(i)$ and $C_r(j)$ be the set of *coreference* links formed by mentions in k_i and r_j:

$$C_k(i) = \{(m_1, m_2) : m_1 \in k_i, m_2 \in k_i, m_1 \neq m_2\}$$
$$C_r(j) = \{(m_1, m_2) : m_1 \in r_j, m_2 \in r_j, m_1 \neq m_2\}$$

As can be seen, a link is an undirected edge between two mentions, and it can be equivalently represented by a pair of mentions. Note that when an entity consists of a single mention, its coreference link set is empty.

Let $N_k(i,j)$ $(i \neq j)$ be key *non-coreference* links formed between mentions in k_i and those in k_j, and let $N_r(i,j)$ $(i \neq j)$ be response *non-coreference* links formed between mentions in r_i and those in r_j, respectively:

$$N_k(i,j) = \{(m_1, m_2) : m_1 \in k_i, m_2 \in k_j\}$$
$$N_r(i,j) = \{(m_1, m_2) : m_1 \in r_i, m_2 \in r_j\}$$

Note that the non-coreference link set is empty when all mentions are in the same entity.

We use the same letter and subscription without the index in parentheses to denote the union of sets, e.g.,

$$C_k = \cup_i C_k(i),\ N_k = \cup_{i \neq j} N_k(i,j)$$
$$C_r = \cup_j C_r(j),\ N_r = \cup_{i \neq j} N_r(i,j)$$

We use $T_k = C_k \cup N_k$ and $T_r = C_r \cup N_r$ to denote the total set of key links and total set of response links, respectively. Clearly, C_k and N_k form a partition of T_k since $C_k \cap N_k = \emptyset$, $T_k = C_k \cup N_k$. Likewise, C_r and N_r form a partition of T_r.

We say that a key link $l_1 \in T_k$ equals a response link $l_2 \in T_r$ if and only if the pair of mentions from which the links are formed are identical. We write $l_1 = l_2$ if two links are equal. It is easy to see that the gold mention assumption—same set of response mentions as the set of key mentions—can be equivalently stated as $T_k = T_r$ (this does not necessarily mean that $C_k = C_r$ or $N_k = N_r$).

We also use $|\cdot|$ to denote the size of a set.

3 Original BLANC

BLANC-gold is adapted from Rand Index (Rand, 1971), a metric for clustering objects. Rand Index is defined as the ratio between the number of correct within-cluster links plus the number of correct cross-cluster links, and the total number of links.

When $T_k = T_r$, Rand Index can be applied directly since coreference resolution reduces to a clustering problem where mentions are partitioned into clusters (entities):

$$\text{Rand Index} = \frac{|C_k \cap C_r| + |N_k \cap N_r|}{\frac{1}{2}\big(|T_k|(|T_k| - 1)\big)} \tag{1}$$

In practice, though, the simple-minded adoption of Rand Index is not satisfactory since the number of non-coreference links often overwhelms that of coreference links (Recasens and Hovy, 2011), or, $|N_k| \gg |C_k|$ and $|N_r| \gg |C_r|$. Rand Index, if used without modification, would not be sensitive to changes of coreference links.

BLANC-gold solves this problem by averaging the F-measure computed over coreference links and the F-measure over non-coreference links. Using the notations in Section 2, the recall, precision, and F-measure on coreference links are:

$$R_c^{(g)} = \frac{|C_k \cap C_r|}{|C_k \cap C_r| + |C_k \cap N_r|} \tag{2}$$

$$P_c^{(g)} = \frac{|C_k \cap C_r|}{|C_r \cap C_k| + |C_r \cap N_k|} \tag{3}$$

$$F_c^{(g)} = \frac{2R_c^{(g)} P_c^{(g)}}{R_c^{(g)} + P_c^{(g)}}; \tag{4}$$

Similarly, the recall, precision, and F-measure on non-coreference links are computed as:

$$R_n^{(g)} = \frac{|N_k \cap N_r|}{|N_k \cap C_r| + |N_k \cap N_r|} \tag{5}$$

$$P_n^{(g)} = \frac{|N_k \cap N_r|}{|N_r \cap C_k| + |N_r \cap N_k|} \tag{6}$$

$$F_n^{(g)} = \frac{2R_n^{(g)} P_n^{(g)}}{R_n^{(g)} + P_n^{(g)}}. \tag{7}$$

Finally, the BLANC-gold metric is the arithmetic average of $F_c^{(g)}$ and $F_n^{(g)}$:

$$\text{BLANC}^{(g)} = \frac{F_c^{(g)} + F_n^{(g)}}{2}. \quad (8)$$

Superscript ^{g} in these equations highlights the fact that they are meant for coreference systems with gold mentions.

Eqn. (8) indicates that BLANC-gold assigns equal weight to $F_c^{(g)}$, the F-measure from coreference links, and $F_n^{(g)}$, the F-measure from non-coreference links. This avoids the problem that $|N_k| \gg |C_k|$ and $|N_r| \gg |C_r|$, should the original Rand Index be used.

In Eqn. (2) - (3) and Eqn. (5) - (6), denominators are written as a sum of disjoint subsets so they can be related to the contingency table in (Recasens and Hovy, 2011). Under the assumption that $T_k = T_r$, it is clear that $C_k = (C_k \cap C_r) \cup (C_k \cap N_r)$, $C_r = (C_k \cap C_r) \cup (N_k \cap C_r)$, and so on.

4 BLANC for Imperfect Response Mentions

Under the assumption that the key and response mention sets are identical (which implies that $T_k = T_r$), Equations (2) to (7) make sense. For example, R_c is the ratio of the number of correct coreference links over the number of key coreference links; P_c is the ratio of the number of correct coreference links over the number of response coreference links, and so on.

However, when response mentions are not identical to key mentions, a key coreference link may not appear in either C_r or N_r, so Equations (2) to (7) cannot be applied directly to systems with imperfect mentions. For instance, if the key entities are {`a,b,c`} {`d,e`}; and the response entities are {`b,c`} {`e,f,g`}, then the key coreference link `(a,b)` is not seen on the response side; similarly, it is possible that a response link does not appear on the key side either: `(c,f)` and `(f,g)` are not in the key in the above example.

To account for missing or spurious links, we observe that

- $C_k \setminus T_r$ are key coreference links missing in the response;
- $N_k \setminus T_r$ are key non-coreference links missing in the response;
- $C_r \setminus T_k$ are response coreference links missing in the key;
- $N_r \setminus T_k$ are response non-coreference links missing in the key,

and we propose to extend the coreference F-measure and non-coreference F-measure as follows. Coreference recall, precision and F-measure are changed to:

$$R_c = \frac{|C_k \cap C_r|}{|C_k \cap C_r| + |C_k \cap N_r| + |C_k \setminus T_r|} \quad (9)$$

$$P_c = \frac{|C_k \cap C_r|}{|C_r \cap C_k| + |C_r \cap N_k| + |C_r \setminus T_k|} \quad (10)$$

$$F_c = \frac{2R_cP_c}{R_c + P_c} \quad (11)$$

Non-coreference recall, precision and F-measure are changed to:

$$R_n = \frac{|N_k \cap N_r|}{|N_k \cap C_r| + |N_k \cap N_r| + |N_k \setminus T_r|} \quad (12)$$

$$P_n = \frac{|N_k \cap N_r|}{|N_r \cap C_k| + |N_r \cap N_k| + |N_r \setminus T_k|} \quad (13)$$

$$F_n = \frac{2R_nP_n}{R_n + P_n}. \quad (14)$$

The proposed BLANC continues to be the arithmetic average of F_c and F_n:

$$\text{BLANC} = \frac{F_c + F_n}{2}. \quad (15)$$

We observe that the definition of the proposed BLANC, Equ. (9)-(14) subsume the BLANC-gold (2) to (7) due to the following proposition:
If $T_k = T_r$, then $BLANC = BLANC^{(g)}$.

Proof. We only need to show that $R_c = R_c^{(g)}$, $P_c = P_c^{(g)}$, $R_n = R_n^{(g)}$, and $P_n = P_n^{(g)}$. We prove the first one (the other proofs are similar and elided due to space limitations). Since $T_k = T_r$ and $C_k \subset T_k$, we have $C_k \subset T_r$; thus $C_k \setminus T_r = \emptyset$, and $|C_k \cap T_r| = 0$. This establishes that $R_c = R_c^{(g)}$.

Indeed, since C_k is a union of three disjoint subsets: $C_k = (C_k \cap C_r) \cup (C_k \cap N_r) \cup (C_k \setminus T_r)$, $R_c^{(g)}$ and R_c can be unified as $\frac{|C_k \cap C_r|}{|C_K|}$. Unification for other component recalls and precisions can be done similarly. So the final definition of BLANC can be succinctly stated as:

$$R_c = \frac{|C_k \cap C_r|}{|C_k|}, \quad P_c = \frac{|C_k \cap C_r|}{|C_r|} \quad (16)$$

$$R_n = \frac{|N_k \cap N_r|}{|N_k|}, \quad P_n = \frac{|N_k \cap N_r|}{|N_r|} \quad (17)$$

$$F_c = \frac{2|C_k \cap C_r|}{|C_k| + |C_r|}, \quad F_n = \frac{2|N_k \cap N_r|}{|N_k| + |N_r|} \quad (18)$$

$$\text{BLANC} = \frac{F_c + F_n}{2} \quad (19)$$

4.1 Boundary Cases

Care has to be taken when counts of the BLANC definition are 0. This can happen when all key (or response) mentions are in one cluster or are all singletons: the former case will lead to $N_k = \emptyset$ (or $N_r = \emptyset$); the latter will lead to $C_k = \emptyset$ (or $C_r = \emptyset$). Observe that as long as $|C_k| + |C_r| > 0$, F_c in (18) is well-defined; as long as $|N_k|+|N_r| > 0$, F_n in (18) is well-defined. So we only need to augment the BLANC definition for the following cases:

(1) If $C_k = C_r = \emptyset$ and $N_k = N_r = \emptyset$, then BLANC $= I(M_k = M_r)$, where $I(\cdot)$ is an indicator function whose value is 1 if its argument is true, and 0 otherwise. M_k and M_r are the key and response mention set. This can happen when a document has no more than one mention and there is no link.

(2) If $C_k = C_r = \emptyset$ and $|N_k| + |N_r| > 0$, then BLANC $= F_n$. This is the case where the key and response side has only entities consisting of singleton mentions. Since there is no coreference link, BLANC reduces to the non-coreference F-measure F_n.

(3) If $N_k = N_r = \emptyset$ and $|C_k| + |C_r| > 0$, then BLANC $= F_c$. This is the case where all mentions in the key and response are in one entity. Since there is no non-coreference link, BLANC reduces to the coreference F-measure F_c.

4.2 Toy Examples

We walk through a few examples and show how BLANC is calculated in detail. In all the examples below, each lower-case letter represents a mention; mentions in an entity are closed in {}; two letters in () represent a link.

Example 1. Key entities are $\{abc\}$ and $\{d\}$; response entities are $\{bc\}$ and $\{de\}$. Obviously,

$$C_k = \{(ab), (bc), (ac)\};$$
$$N_k = \{(ad), (bd), (cd)\};$$
$$C_r = \{(bc), (de)\};$$
$$N_r = \{(bd), (be), (cd), (ce)\}.$$

Therefore, $C_k \cap C_r = \{(bc)\}$, $N_k \cap N_r = \{(bd), (cd)\}$, and $R_c = \frac{1}{3}$, $P_c = \frac{1}{2}$, $F_c = \frac{2}{5}$; $R_n = \frac{2}{3}$, $P_n = \frac{2}{4}$, $F_n = \frac{4}{7}$. Finally, BLANC $= \frac{17}{35}$.

Example 2. Key entity is $\{a\}$; response entity is $\{b\}$. This is boundary case (1): BLANC $= 0$.

Example 3. Key entities are $\{a\}\{b\}\{c\}$; response entities are $\{a\}\{b\}\{d\}$. This is boundary case (2): there are no coreference links. Since

$$N_k = \{(ab), (bc), (ca)\},$$
$$N_r = \{(ab), (bd), (ad)\},$$

we have

$N_k \cap N_r = \{(ab)\}$, and $R_n = \frac{1}{3}$, $P_n = \frac{1}{3}$.

So BLANC $= F_n = \frac{1}{3}$.

Example 4. Key entity is $\{abc\}$; response entity is $\{bc\}$. This is boundary case (3): there are no non-coreference links. Since

$C_k = \{(ab), (bc), (ca)\}$, and $C_r = \{(bc)\}$,

we have

$C_k \cap C_r = \{(bc)\}$, and $R_c = \frac{1}{3}$, $P_c = 1$,

So BLANC $= F_c = \frac{2}{4} = \frac{1}{2}$.

Participant	R	P	BLANC
lee	50.23	49.28	48.84
sapena	40.68	49.05	44.47
nugues	47.83	44.22	45.95
chang	44.71	47.48	45.49
stoyanov	49.37	29.80	34.58
santos	46.74	37.33	41.33
song	36.88	39.69	30.92
sobha	35.42	39.56	36.31
yang	47.95	29.12	36.09
charton	42.32	31.54	35.65
hao	45.41	32.75	36.98
zhou	29.93	45.58	34.95
kobdani	32.29	33.01	32.57
xinxin	36.83	34.39	35.02
kummerfeld	34.84	29.53	30.98
zhang	30.10	43.96	35.71
zhekova	26.40	15.32	15.37
irwin	3.62	28.28	6.28

Table 1: The proposed BLANC scores of the CoNLL-2011 shared task participants.

5 Results

5.1 CoNLL-2011/12

We have updated the publicly available CoNLL coreference scorer[1] with the proposed BLANC, and used it to compute the proposed BLANC scores for all the CoNLL 2011 (Pradhan et al., 2011) and 2012 (Pradhan et al., 2012) participants in the official track, where participants had to automatically predict the mentions. Tables 1 and 2 report the updated results.[2]

5.2 Correlation with Other Measures

Figure 1 shows how the proposed BLANC measure works when compared with existing metrics such as MUC, B-cubed and CEAF, using the BLANC and F1 scores. The proposed BLANC is highly positively correlated with the

[1] http://code.google.com/p/reference-coreference-scorers

[2] The order is kept the same as in Pradhan et al. (2011) and Pradhan et al. (2012) for easy comparison.

Participant	R	P	BLANC
Language: Arabic			
fernandes	33.43	44.66	37.99
bjorkelund	32.65	45.47	37.93
uryupina	31.62	35.26	33.02
stamborg	32.59	36.92	34.50
chen	31.81	31.52	30.82
zhekova	11.04	62.58	18.51
li	4.60	56.63	8.42
Language: English			
fernandes	54.91	63.66	58.75
martschat	52.00	58.84	55.04
bjorkelund	52.01	59.55	55.42
chang	52.85	55.03	53.86
chen	50.52	56.82	52.87
chunyang	51.19	55.47	52.65
stamborg	54.39	54.88	54.42
yuan	50.58	54.29	52.11
xu	45.99	54.59	46.47
shou	49.55	52.46	50.44
uryupina	44.15	48.89	46.04
songyang	40.60	50.85	45.10
zhekova	41.46	33.13	34.80
xinxin	44.39	32.79	36.54
li	25.17	52.96	31.85
Language: Chinese			
chen	48.45	62.44	54.10
yuan	53.15	40.75	43.20
bjorkelund	47.58	45.93	44.22
xu	44.11	36.45	38.45
fernandes	42.36	61.72	49.63
stamborg	39.60	55.12	45.89
uryupina	33.44	56.01	41.88
martschat	27.24	62.33	37.89
chunyang	37.43	36.18	36.77
xinxin	36.46	39.79	37.85
li	21.61	62.94	30.37
chang	18.74	40.76	25.68
zhekova	21.50	37.18	22.89

Table 2: The proposed BLANC scores of the CoNLL-2012 shared task participants.

	R	P	F1
MUC	0.975	0.844	0.935
B-cubed	0.981	0.942	0.966
CEAF-m	0.941	0.923	0.966
CEAF-e	0.797	0.781	0.919

Table 3: Pearson's r correlation coefficients between the proposed BLANC and the other coreference measures based on the CoNLL 2011/2012 results. All p-values are significant at < 0.001.

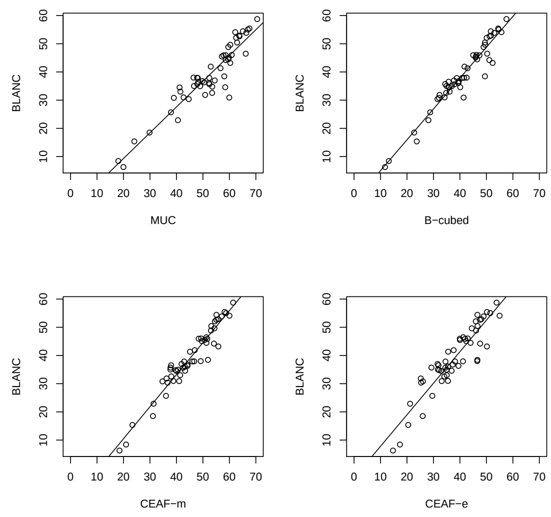

Figure 1: Correlation plot between the proposed BLANC and the other measures based on the CoNLL 2011/2012 results. All values are F1 scores.

other measures along R, P and F1 (Table 3), showing that BLANC is able to capture most entity-based similarities measured by B-cubed and CEAF. However, the CoNLL data sets come from OntoNotes (Hovy et al., 2006), where singleton entities are not annotated, and BLANC has a wider dynamic range on data sets with singletons (Recasens and Hovy, 2011). So the correlations will likely be lower on data sets with singleton entities.

6 Conclusion

The original BLANC-gold (Recasens and Hovy, 2011) requires that system mentions be identical to gold mentions, which limits the metric's utility since detected system mentions often have missing key mentions or spurious mentions. The proposed BLANC is free from this assumption, and we have shown that it subsumes the original BLANC-gold. Since BLANC works on imperfect system mentions, we have used it to score the CoNLL 2011 and 2012 coreference systems. The BLANC scores show strong correlation with existing metrics, especially B-cubed and CEAF-m.

Acknowledgments

We would like to thank the three anonymous reviewers for their invaluable suggestions for improving the paper. This work was partially supported by grants R01LM10090 from the National Library of Medicine.

References

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *Proceedings of the Linguistic Coreference Workshop at The First International Conference on Language Resources and Evaluation (LREC'98)*, pages 563–566.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, New York City, USA, June. Association for Computational Linguistics.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proc. of Human Language Technology (HLT)/Empirical Methods in Natural Language Processing (EMNLP)*.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. CoNLL-2011 shared task: Modeling unrestricted coreference in OntoNotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–27, Portland, Oregon, USA, June. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, pages 1–40, Jeju Island, Korea, July. Association for Computational Linguistics.

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

M. Recasens and E. Hovy. 2011. BLANC: Implementing the Rand index for coreference evaluation. *Natural Language Engineering*, 17:485–510, 10.

Veselin Stoyanov, Nathan Gilbert, Claire Cardie, and Ellen Riloff. 2009. Conundrums in noun phrase coreference resolution: Making sense of the state-of-the-art. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2*, ACL '09, pages 656–664, Stroudsburg, PA, USA. Association for Computational Linguistics.

M. Vilain, J. Burger, J. Aberdeen, D. Connolly, , and L. Hirschman. 1995. A model-theoretic coreference scoring scheme. In *In Proc. of MUC6*, pages 45–52.