Scoring Coreference Partitions of Predicted Mentions: A Reference Implementation

Sameer Pradhan[1], **Xiaoqiang Luo**[2], **Marta Recasens**[3], **Eduard Hovy**[4], **Vincent Ng**[5] **and Michael Strube**[6]

[1]Harvard Medical School, Boston, MA, [2]Google Inc., New York, NY
[3]Google Inc., Mountain View, CA, [4]Carnegie Mellon University, Pittsburgh, PA
[5]HLTRI, University of Texas at Dallas, Richardson, TX, [6]HITS, Heidelberg, Germany

sameer.pradhan@childrens.harvard.edu, {xql,recasens}@google.com,
hovy@cmu.edu, vince@hlt.utdallas.edu, michael.strube@h-its.org

Abstract

The definitions of two coreference scoring metrics—B^3 and CEAF—are underspecified with respect to *predicted*, as opposed to *key* (or *gold*) mentions. Several variations have been proposed that manipulate either, or both, the key and predicted mentions in order to get a one-to-one mapping. On the other hand, the metric BLANC was, until recently, limited to scoring partitions of key mentions. In this paper, we (i) argue that mention manipulation for scoring predicted mentions is unnecessary, and potentially harmful as it could produce unintuitive results; (ii) illustrate the application of all these measures to scoring predicted mentions; (iii) make available an open-source, thoroughly-tested reference implementation of the main coreference evaluation measures; and (iv) rescore the results of the CoNLL-2011/2012 shared task systems with this implementation. This will help the community accurately measure and compare new end-to-end coreference resolution algorithms.

1 Introduction

Coreference resolution is a key task in natural language processing (Jurafsky and Martin, 2008) aiming to detect the referential expressions (*mentions*) in a text that point to the same entity. Roughly over the past two decades, research in coreference (for the English language) had been plagued by individually crafted evaluations based on two central corpora—MUC (Hirschman and Chinchor, 1997; Chinchor and Sundheim, 2003; Chinchor, 2001) and ACE (Doddington et al., 2004). Experimental parameters ranged from using perfect (*gold*, or *key*) mentions as input for purely testing the quality of the entity linking algorithm, to an end-to-end evaluation where *predicted mentions* are used. Given the range of evaluation parameters and disparity between the annotation standards for the two corpora, it was very hard to grasp the state of the art for the task of coreference. This has been expounded in Stoyanov et al. (2009). The activity in this subfield of NLP can be gauged by: (i) the continual addition of corpora manually annotated for coreference—The OntoNotes corpus (Pradhan et al., 2007; Weischedel et al., 2011) in the general domain, as well as the i2b2 (Uzuner et al., 2012) and THYME (Styler et al., 2014) corpora in the clinical domain would be a few examples of such emerging corpora; and (ii) ongoing proposals for refining the existing metrics to make them more informative (Holen, 2013; Chen and Ng, 2013).

The CoNLL-2011/2012 shared tasks on coreference resolution using the OntoNotes corpus (Pradhan et al., 2011; Pradhan et al., 2012) were an attempt to standardize the evaluation settings by providing a benchmark annotated corpus, scorer, and state-of-the-art system results that would allow future systems to compare against them. Following the timely emphasis on end-to-end evaluation, the official track used predicted mentions and measured performance using five coreference measures: MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), $CEAF_e$ (Luo, 2005), $CEAF_m$ (Luo, 2005), and BLANC (Recasens and Hovy, 2011). The arithmetic mean of the first three was the task's final score.

An unfortunate setback to these evaluations had its root in three issues: (i) the multiple variations of two of the scoring metrics—B^3 and CEAF—used by the community to handle predicted mentions; (ii) a buggy implementation of the Cai and Strube (2010) proposal that tried to reconcile these variations; and (iii) the erroneous computation of

the BLANC metric for partitions of predicted mentions. Different interpretations as to how to compute B^3 and CEAF scores for coreference systems when predicted mentions do not perfectly align with key mentions—which is usually the case—led to variations of these metrics that manipulate the gold standard and system output in order to get a one-to-one mention mapping (Stoyanov et al., 2009; Cai and Strube, 2010). Some of these variations arguably produce rather unintuitive results, while others are not faithful to the original measures.

In this paper, we address the issues in scoring coreference partitions of predicted mentions. Specifically, we justify our decision to go back to the original scoring algorithms by arguing that manipulation of key or predicted mentions is unnecessary and could in fact produce unintuitive results. We demonstrate the use of our recent extension of BLANC that can seamlessly handle predicted mentions (Luo et al., 2014). We make available an open-source, thoroughly-tested reference implementation of the main coreference evaluation measures that do not involve mention manipulation and is faithful to the original intentions of the proposers of these metrics. We republish the CoNLL-2011/2012 results based on this scorer, so that future systems can use it for evaluation and have the CoNLL results available for comparison.

The rest of the paper is organized as follows. Section 2 provides an overview of the variations of the existing measures. We present our newly updated coreference scoring package in Section 3 together with the rescored CoNLL-2011/2012 outputs. Section 4 walks through a scoring example for all the measures, and we conclude in Section 5.

2 Variations of Scoring Measures

Two commonly used coreference scoring metrics —B^3 and CEAF—are underspecified in their application for scoring *predicted*, as opposed to *key* mentions. The examples in the papers describing these metrics assume perfect mentions where predicted mentions are the same set of mentions as key mentions. The lack of accompanying reference implementation for these metrics by its proposers made it harder to fill the gaps in the specification. Subsequently, different interpretations of how one can evaluate coreference systems when predicted mentions do not perfectly align with key mentions led to variations of these metrics that manipulate the gold and/or predicted mentions (Stoyanov et al., 2009; Cai and Strube, 2010). All these variations attempted to generate a one-to-one mapping between the key and predicted mentions, assuming that the original measures cannot be applied to predicted mentions. Below we first provide an overview of these variations and then discuss the unnecessity of this assumption.

Coining the term *twinless mentions* for those mentions that are either spurious or missing from the predicted mention set, Stoyanov et al. (2009) proposed two variations to B^3 — B^3_{all} and B^3_0—to handle them. In the first variation, all predicted twinless mentions are retained, whereas the latter discards them and penalizes recall for twinless predicted mentions. Rahman and Ng (2009) proposed another variation by removing "all and only those twinless system mentions that are singletons before applying B^3 and CEAF." Following upon this line of research, Cai and Strube (2010) proposed a unified solution for both B^3 and CEAF_m, leaving the question of handling CEAF_e as future work because "it produces unintuitive results." The essence of their solution involves manipulating twinless key and predicted mentions by adding them either from the predicted partition to the key partition or vice versa, depending on whether one is computing precision or recall. The Cai and Strube (2010) variation was used by the CoNLL-2011/2012 shared tasks on coreference resolution using the OntoNotes corpus, and by the i2b2 2011 shared task on coreference resolution using an assortment of clinical notes corpora (Uzuner et al., 2012).[1] It was later identified by Recasens et al. (2013) that there was a bug in the implementation of this variation in the scorer used for the CoNLL-2011/2012 tasks. We have not tested the correctness of this variation in the scoring package used for the i2b2 shared task.

However, it turns out that the CEAF metric (Luo, 2005) was always intended to work seamlessly on predicted mentions, and so has been the case with the B^3 metric.[2] In a latter paper, Rahman and Ng (2011) correctly state that "CEAF can compare partitions with twinless mentions without any modification." We will look at this further in Section 4.3.

We argue that manipulations of key and response mentions/entities, as is done in the existing B^3 variations, not only confound the evaluation process, but are also subject to abuse and can seriously jeopardize the fidelity of the evalu-

[1] Personal communication with Andreea Bodnari, and contents of the i2b2 scorer code.

[2] Personal communication with Breck Baldwin.

ation. Given space constraints we use an example worked out in Cai and Strube (2010). Let the key contain an entity with mentions $\{a, b, c\}$ and the prediction contain an entity with mentions $\{a, b, d\}$. As detailed in Cai and Strube (2010, p. 29-30, Tables 1–3), B^3_0 assigns a perfect precision of 1.00 which is unintuitive as the system has wrongly predicted a mention d as belonging to the entity. For the same prediction, B^3_{all} assigns a precision of 0.556. But, if the prediction contains two entities $\{a, b, d\}$ and $\{c\}$ (i.e., the mention c is added as a spurious singleton), then B^3_{all} precision increases to 0.667 which is counter-intuitive as it does not penalize the fact that c is erroneously placed in its own entity. The version illustrated in Section 4.2, which is devoid of any mention manipulations, gives a precision of 0.444 in the first scenario and the precision drops to 0.333 in the second scenario with the addition of a spurious singleton entity $\{c\}$. This is a more intuitive behavior.

Contrary to both B^3 and CEAF, the BLANC measure (Recasens and Hovy, 2011) was never designed to handle predicted mentions. However, the implementation used for the SemEval-2010 shared task as well as the one for the CoNLL-2011/2012 shared tasks accepted predicted mentions as input, producing undefined results. In Luo et al. (2014) we have extended the BLANC metric to deal with predicted mentions

3 Reference Implementation

Given the potential unintuitive outcomes of mention manipulation and the misunderstanding that the original measures could not handle twinless predicted mentions (Section 2), we redesigned the CoNLL scorer. The new implementation:

- is faithful to the original measures;
- removes any prior mention manipulation, which might depend on specific annotation guidelines among other problems;
- has been thoroughly tested to ensure that it gives the expected results according to the original papers, and all test cases are included as part of the release;
- is free of the reported bugs that the CoNLL scorer (v4) suffered (Recasens et al., 2013);
- includes the extension of BLANC to handle predicted mentions (Luo et al., 2014).

This is the open source scoring package[3] that we present as a reference implementation for the

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

SYSTEM	MD	MUC	B^3	CEAF		BLANC	CONLL AVERAGE
				m	e		
	F_1	F^1_1	F^2_1	F_1	F^3_1		
CoNLL-2011; English							
lee	70.7	59.6	48.9	53.0	46.1	48.8	51.5
sapena	68.4	59.5	46.5	51.3	44.0	44.5	50.0
nugues	69.0	58.6	45.0	48.4	40.0	46.0	47.9
chang	64.9	57.2	46.0	50.7	40.0	45.5	47.7
stoyanov	67.8	58.4	40.1	43.3	36.9	34.6	45.1
santos	65.5	56.7	42.9	45.1	35.6	41.3	45.0
song	67.3	60.0	41.4	41.0	33.1	30.9	44.8
sobha	64.8	50.5	39.5	44.2	39.4	36.3	43.1
yang	63.9	52.3	39.4	43.2	35.5	36.1	42.4
charton	64.3	52.5	38.0	42.6	34.5	35.7	41.6
hao	64.3	54.5	37.7	41.9	31.6	37.0	41.3
zhou	62.3	49.0	37.0	40.6	35.0	35.0	40.3
kobdani	61.0	53.5	34.8	38.1	34.1	32.6	38.7
xinxin	61.9	46.6	34.9	37.7	31.7	35.0	37.7
kummerfeld	62.7	42.7	34.2	38.8	35.5	31.0	37.5
zhang	61.1	47.9	34.4	37.8	29.2	35.7	37.2
zhekova	48.3	24.1	23.7	23.4	20.5	15.4	22.8
irwin	26.7	20.0	11.7	18.5	14.7	6.3	15.5
CoNLL-2012; English							
fernandes	77.7	70.5	57.6	61.4	53.9	58.8	60.7
martschat	75.2	67.0	54.6	58.8	51.5	55.0	57.7
bjorkelund	75.4	67.6	54.5	58.2	50.2	55.4	57.4
chang	74.3	66.4	53.0	57.1	48.9	53.9	56.1
chen	73.8	63.7	51.8	55.8	48.1	52.9	54.5
chunyang	73.7	63.8	51.2	55.1	47.6	52.7	54.2
stamborg	73.9	65.1	51.7	55.1	46.6	54.4	54.2
yuan	72.5	62.6	50.1	54.5	46.0	52.1	52.9
xu	72.0	66.2	50.3	51.3	41.3	46.5	52.6
shou	73.7	62.9	49.4	53.2	46.7	50.4	53.0
uryupina	70.9	60.9	46.2	49.3	42.9	46.0	50.0
songyang	68.8	59.8	45.9	49.6	42.4	45.1	49.4
zhekova	67.1	53.5	35.7	39.7	32.2	34.8	40.5
xinxin	62.8	48.3	35.7	38.0	31.9	36.5	38.6
li	59.9	50.8	32.3	36.3	25.2	31.9	36.1
CoNLL-2012; Chinese							
chen	71.6	62.2	55.7	60.0	55.0	54.1	57.6
yuan	68.2	60.3	52.4	55.8	50.2	43.2	54.3
bjorkelund	66.4	58.6	51.1	54.2	47.6	44.2	52.5
xu	65.2	58.1	49.5	51.9	46.6	38.5	51.4
fernandes	66.1	60.3	49.6	54.4	44.5	49.6	51.5
stamborg	64.0	57.8	47.4	51.6	41.9	45.9	49.0
uryupina	59.0	53.0	41.7	46.9	37.6	41.9	44.1
martschat	58.6	52.4	40.8	46.0	38.2	37.9	43.8
chunyang	61.6	49.8	39.6	44.2	37.3	36.8	42.2
xinxin	55.9	48.1	38.8	42.9	34.5	37.9	40.5
li	51.5	44.7	31.5	36.7	25.3	30.4	33.8
chang	47.6	37.9	28.8	36.1	29.6	25.7	32.1
zhekova	47.3	40.6	28.1	31.4	21.2	22.9	30.0
CoNLL-2012; Arabic							
fernandes	64.8	46.5	42.5	49.2	46.5	38.0	45.2
bjorkelund	60.6	47.8	41.6	46.7	41.2	37.9	43.5
uryupina	55.4	41.5	36.1	41.4	35.0	33.0	37.5
stamborg	59.5	41.2	35.9	40.0	32.9	34.5	36.7
chen	59.8	39.0	32.1	34.7	26.0	30.8	32.4
zhekova	41.0	29.9	22.7	31.1	25.9	18.5	26.2
li	29.7	18.1	13.1	21.0	17.3	8.4	16.2

Table 1: Performance on the **official, closed** track in percentages using all predicted information for the CoNLL-2011 and 2012 shared tasks.

community to use. It is written in perl and stems from the scorer that was initially used for the SemEval-2010 shared task (Recasens et al., 2010) and later modified for the CoNLL-2011/2012 shared tasks.[4]

Partitioning detected mentions into entities (or equivalence classes) typically comprises two distinct tasks: (i) mention detection; and (ii) coreference resolution. A typical two-step coreference algorithm uses mentions generated by the best

[4] We would like to thank Emili Sapena for writing the first version of the scoring package.

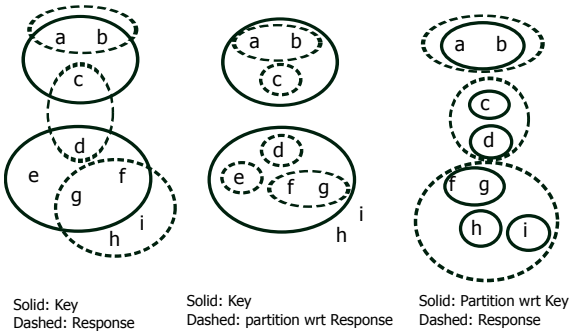

Figure 1: Example key and response entities along with the partitions for computing the MUC score.

possible mention detection algorithm as input to the coreference algorithm. Therefore, ideally one would want to score the two steps independently of each other. A peculiarity of the OntoNotes corpus is that singleton referential mentions are not annotated, thereby preventing the computation of a mention detection score independently of the coreference resolution score. In corpora where all referential mentions (including singletons) are annotated, the mention detection score generated by this implementation is independent of the coreference resolution score.

We used this reference implementation to rescore the CoNLL-2011/2012 system outputs for the official task to enable future comparisons with these benchmarks. The new CoNLL-2011/2012 results are in Table 1. We found that the overall system ranking remained largely unchanged for both shared tasks, except for some of the lower ranking systems that changed one or two places. However, there was a considerable drop in the magnitude of all B^3 scores owing to the combination of two things: (i) mention manipulation, as proposed by Cai and Strube (2010), adds singletons to account for *twinless* mentions; and (ii) the B^3 metric allows an entity to be used more than once as pointed out by Luo (2005). This resulted in a drop in the CoNLL averages (B^3 is one of the three measures that make the average).

4 An Illustrative Example

This section walks through the process of computing each of the commonly used metrics for an example where the set of predicted mentions has some missing key mentions and some spurious mentions. While the mathematical formulae for these metrics can be found in the original papers (Vilain et al., 1995; Bagga and Baldwin, 1998; Luo, 2005), many misunderstandings discussed in Section 2 are due to the fact that these papers lack an example showing how a metric is computed on predicted mentions. A concrete example goes a long way to prevent similar misunderstandings in the future. The example is adapted from Vilain et al. (1995) with some slight modifications so that the total number of mentions in the key is different from the number of mentions in the prediction. The key (K) contains two entities with mentions $\{a, b, c\}$ and $\{d, e, f, g\}$ and the response (R) contains three entities with mentions $\{a, b\}$; $\{c, d\}$ and $\{f, g, h, i\}$:

$$K = \overbrace{\{a, b, c\}}^{K_1}\ \overbrace{\{d, e, f, g\}}^{K_2} \quad (1)$$

$$R = \overbrace{\{a, b\}}^{R_1}\ \overbrace{\{c, d\}}^{R_2}\ \overbrace{\{f, g, h, i\}}^{R_3}. \quad (2)$$

Mention e is missing from the response, and mentions h and i are spurious in the response. The following sections use R to denote recall and P for precision.

4.1 MUC

The main step in the MUC scoring is creating the partitions with respect to the key and response respectively, as shown in Figure 1. Once we have the partitions, then we compute the MUC score by:

$$\begin{aligned} R &= \frac{\sum_{i=1}^{N_k}(|K_i| - |p(K_i)|)}{\sum_{i=1}^{N_k}(|K_i| - 1)} \\ &= \frac{(3-2)+(4-3)}{(3-1)+(4-1)} = 0.40 \\ P &= \frac{\sum_{i=1}^{N_r}(|R_i| - |p'(R_i)|)}{\sum_{i=1}^{N_r}(|R_i| - 1)} \\ &= \frac{(2-1)+(2-2)+(4-3)}{(2-1)+(2-1)+(4-1)} = 0.40, \end{aligned}$$

where K_i is the i^{th} key entity and $p(K_i)$ is the set of partitions created by intersecting K_i with response entities (cf. the middle sub-figure in Figure 1); R_i is the i^{th} response entity and $p'(R_i)$ is the set of partitions created by intersecting R_i with key entities (cf. the right-most sub-figure in Figure 1); and N_k and N_r are the number of key and response entities, respectively.

The MUC F_1 score in this case is 0.40.

4.2 B^3

For computing B^3 recall, each key mention is assigned a credit equal to the ratio of the number of correct mentions in the predicted entity *containing* the key mention to the size of the key entity to which the mention belongs, and the recall is just the sum of credits over all key mentions normalized over the number of key mentions. B^3 preci-

sion is computed similarly, except switching the role of key and response. Applied to the example:

$$R = \frac{\sum_{i=1}^{N_k}\sum_{j=1}^{N_r}\frac{|K_i \cap R_j|^2}{|K_i|}}{\sum_{i=1}^{N_k}|K_i|}$$
$$= \frac{1}{7} \times (\frac{2^2}{3} + \frac{1^2}{3} + \frac{1^2}{4} + \frac{2^2}{4}) = \frac{1}{7} \times \frac{35}{12} \approx 0.42$$
$$P = \frac{\sum_{i=1}^{N_k}\sum_{j=1}^{N_r}\frac{|K_i \cap R_j|^2}{|R_j|}}{\sum_{i=1}^{N_r}|R_j|}$$
$$= \frac{1}{8} \times (\frac{2^2}{2} + \frac{1^2}{2} + \frac{1^2}{2} + \frac{2^2}{4}) = \frac{1}{8} \times \frac{4}{1} = 0.50$$

Note that terms with 0 value are omitted. The B^3 F_1 score is 0.46.

4.3 CEAF

The first step in the CEAF computation is getting the best scoring alignment between the key and response entities. In this case the alignment is straightforward. Entity R_1 aligns with K_1 and R_3 aligns with K_2. R_2 remains unaligned.

4.3.1 CEAF_m

CEAF_m recall is the number of aligned mentions divided by the number of key mentions, and precision is the number of aligned mentions divided by the number of response mentions:

$$R = \frac{|K_1 \cap R_1| + |K_2 \cap R_3|}{|K_1| + |K_2|} = \frac{(2+2)}{(3+4)} \approx 0.57$$
$$P = \frac{|K_1 \cap R_1| + |K_2 \cap R_3|}{|R_1| + |R_2| + |R_3|} = \frac{(2+2)}{(2+2+4)} = 0.50$$

The CEAF_m F_1 score is 0.53.

4.3.2 CEAF_e

We use the same notation as in Luo (2005): $\phi_4(K_i, R_j)$ to denote the similarity between a key entity K_i and a response entity R_j. $\phi_4(K_i, R_j)$ is defined as:

$$\phi_4(K_i, R_j) = \frac{2 \times |K_i \cap R_j|}{|K_i| + |R_j|}.$$

CEAF_e recall and precision, when applied to this example, are:

$$R = \frac{\phi_4(K_1, R_1) + \phi_4(K_2, R_3)}{N_k} = \frac{\frac{(2\times2)}{(3+2)} + \frac{(2\times2)}{(4+4)}}{2} = 0.65$$
$$P = \frac{\phi_4(K_1, R_1) + \phi_4(K_2, R_3)}{N_r} = \frac{\frac{(2\times2)}{(3+2)} + \frac{(2\times2)}{(4+4)}}{3} \approx 0.43$$

The CEAF_e F_1 score is 0.52.

4.4 BLANC

The BLANC metric illustrated here is the one in our implementation which extends the original BLANC (Recasens and Hovy, 2011) to predicted mentions (Luo et al., 2014).

Let C_k and C_r be the set of coreference links in the key and response respectively, and N_k and N_r be the set of non-coreference links in the key and response respectively. A link between a mention pair m and n is denoted by mn; then for the example in Figure 1, we have

$$C_k = \{ab, ac, bc, de, df, dg, ef, eg, fg\}$$
$$N_k = \{ad, ae, af, ag, bd, be, bf, bg, cd, ce, cf, cg\}$$
$$C_r = \{ab, cd, fg, fh, fi, gh, gi, hi\}$$
$$N_r = \{ac, ad, af, ag, ah, ai, bc, bd, bf, bg, bh, bi, cf, cg, ch, ci, df, dg, dh, di\}.$$

Recall and precision for coreference links are:

$$R_c = \frac{|C_k \cap C_r|}{|C_k|} = \frac{2}{9} \approx 0.22$$
$$P_c = \frac{|C_k \cap C_r|}{|C_r|} = \frac{2}{8} = 0.25$$

and the coreference F-measure, $F_c \approx 0.23$. Similarly, recall and precision for non-coreference links are:

$$R_n = \frac{|N_k \cap N_r|}{|N_k|} = \frac{8}{12} \approx 0.67$$
$$P_n = \frac{|N_k \cap N_r|}{|N_r|} = \frac{8}{20} = 0.40,$$

and the non-coreference F-measure, $F_n = 0.50$. So the BLANC score is $\frac{F_c+F_n}{2} \approx 0.36$.

5 Conclusion

We have cleared several misunderstandings about coreference evaluation metrics, especially when a response contains imperfect predicted mentions, and have argued against mention manipulations during coreference evaluation. These misunderstandings are caused partially by the lack of illustrative examples to show how a metric is computed on predicted mentions not aligned perfectly with key mentions. Therefore, we provide detailed steps for computing all four metrics on a representative example. Furthermore, we have a reference implementation of these metrics that has been rigorously tested and has been made available to the public as open source software. We reported new scores on the CoNLL 2011 and 2012 data sets, which can serve as the benchmarks for future research work.

Acknowledgments

This work was partially supported by grants R01LM10090 from the National Library of Medicine and IIS-1219142 from the National Science Foundation.

References

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *Proceedings of LREC*, pages 563–566.

Jie Cai and Michael Strube. 2010. Evaluation metrics for end-to-end coreference resolution systems. In *Proceedings of SIGDIAL*, pages 28–36.

Chen Chen and Vincent Ng. 2013. Linguistically aware coreference evaluation metrics. In *Proceedings of the Sixth IJCNLP*, pages 1366–1374, Nagoya, Japan, October.

Nancy Chinchor and Beth Sundheim. 2003. Message understanding conference (MUC) 6. In *LDC2003T13*.

Nancy Chinchor. 2001. Message understanding conference (MUC) 7. In *LDC2001T02*.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ACE) program-tasks, data, and evaluation. In *Proceedings of LREC*.

Lynette Hirschman and Nancy Chinchor. 1997. Coreference task definition (v3.0, 13 jul 97). In *Proceedings of the 7th Message Understanding Conference*.

Gordana Ilic Holen. 2013. Critical reflections on evaluation practices in coreference resolution. In *Proceedings of the NAACL-HLT Student Research Workshop*, pages 1–7, Atlanta, Georgia, June.

Daniel Jurafsky and James H. Martin. 2008. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall. Second Edition.

Xiaoqiang Luo, Sameer Pradhan, Marta Recasens, and Eduard Hovy. 2014. An extension of BLANC to system mentions. In *Proceedings of ACL*, Baltimore, Maryland, June.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of HLT-EMNLP*, pages 25–32.

Sameer Pradhan, Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2007. OntoNotes: A Unified Relational Semantic Representation. *International Journal of Semantic Computing*, 1(4):405–419.

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 CoNLL: Shared Task*, pages 1–27.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Proceedings of CoNLL: Shared Task*, pages 1–40.

Altaf Rahman and Vincent Ng. 2009. Supervised models for coreference resolution. In *Proceedings of EMNLP*, pages 968–977.

Altaf Rahman and Vincent Ng. 2011. Coreference resolution with world knowledge. In *Proceedings of ACL*, pages 814–824.

Marta Recasens and Eduard Hovy. 2011. BLANC: Implementing the Rand Index for coreference evaluation. *Natural Language Engineering*, 17(4):485–510.

Marta Recasens, Lluís Màrquez, Emili Sapena, M. Antònia Martí, Mariona Taulé, Véronique Hoste, Massimo Poesio, and Yannick Versley. 2010. Semeval-2010 task 1: Coreference resolution in multiple languages. In *Proceedings of SemEval*, pages 1–8.

Marta Recasens, Marie-Catherine de Marneffe, and Chris Potts. 2013. The life and death of discourse entities: Identifying singleton mentions. In *Proceedings of NAACL-HLT*, pages 627–633.

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 ACL-IJCNLP*, pages 656–664.

William F. Styler, Steven Bethard an Sean Finan, Martha Palmer, Sameer Pradhan, Piet C de Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova, and James Pustejovsky. 2014. Temporal annotation in the clinical domain. *Transactions of Computational Linguistics*, 2(April):143–154.

Ozlem Uzuner, Andreea Bodnari, Shuying Shen, Tyler Forbush, John Pestian, and Brett R South. 2012. Evaluating the state of the art in coreference resolution for electronic medical records. *Journal of American Medical Informatics Association*, 19(5), September.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model theoretic coreference scoring scheme. In *Proceedings of the 6th Message Understanding Conference*, pages 45–52.

Ralph Weischedel, Eduard Hovy, Mitchell Marcus, Martha Palmer, Robert Belvin, Sameer Pradhan, Lance Ramshaw, and Nianwen Xue. 2011. OntoNotes: A large training corpus for enhanced processing. In Joseph Olive, Caitlin Christianson, and John McCary, editors, *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*. Springer.