Critical Reflections on Evaluation Practices in Coreference Resolution

Gordana Ilić Holen
Department of Informatics
University of Oslo
Norway
gordanil@ifi.uio.no

Abstract

In this paper we revisit the task of quantitative evaluation of coreference resolution systems. We review the most commonly used metrics (MUC, B^3, CEAF and BLANC) on the basis of their evaluation of coreference resolution in five texts from the OntoNotes corpus. We examine both the correlation between the metrics and the degree to which our human judgement of coreference resolution agrees with the metrics. In conclusion we claim that loss of information value is an essential factor, insufficiently adressed in current metrics, in human perception of the degree of success or failure of coreference resolution. We thus conjecture that including a layer of mention information weight could improve both the coreference resolution and its evaluation.

1 Introduction and motivation

Coreference resolution (CR) is the task of linking together multiple expressions of a given entity (Yang et al., 2003). The field has experienced a surge of interest with several shared tasks in recent years: SemEval 2010 (Recasens et al., 2010), CoNLL 2011 (Pradhan et al., 2011) and CoNLL 2012 (Pradhan et al., 2012). However the field has from the very start been riddled with problems related to the scoring and comparison of CR systems. Currently there are five metrics in wider use: MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), the two CEAF metrics (Luo, 2005) and BLANC (Recasens and Hovy, 2011). As there is no global agreement on which metrics are the most appropriate, the above-mentioned shared tasks have used a combination of several metrics to evaluate the contenders. Although coreference resolution is a subproblem of natural language understanding, coreference resolution evaluation metrics have predominately been discussed in terms of abstract entities and hypothetical system errors. In our view, it is of utmost importance to observe actual texts and actual system errors.

2 Background: The metrics

In this section, we will present the five metrics in the usual terms of precision, recall and F-score. We follow the predominant practice and use the term *mention* for individual referring expressions, and *entity* for sets of mentions that refer to the same object (Luo et al., 2004). We use the term *key entity (K)* for gold entities, and *response entity (R)* for entities which were produced by the CR system.

2.1 Link-based: MUC and BLANC

The MUC metric (Vilain et al., 1995) is based on comparing the number of links in the key entity ($|K| - 1$) to the number of links missing from the response entity, routinely calculated as the number of partitions of the key entity $|p(K)|$ minus one, so $Recall = \frac{(|K|-1)-(|p(K)|-1)}{|K|-1} = \frac{|K|-|p(K)|}{|K|-1}$. For the whole document, recalls for entities are simply added: $Recall = \frac{\sum |K_i|-|p(K_i)|}{\sum (|K_i|-1)}$ In calculating precision, the case is inverted: The base entity is now the response, and the question posed is how many missing links have to be added to the key partitions to form the response entity.

BLANC (Recasens and Hovy, 2011) is a variant of the Rand index (Rand, 1971) adapted for the task

Proceedings of the NAACL HLT 2013 Student Research Workshop, pages 1–7,
Atlanta, Georgia, 13 June 2013.

of coreference resolution. The BLANC metric makes use of both coreferent and non-coreferent links, correct and incorrect. The final precision, recall and F-score are the average of the P, R and F-score of corresponding coreferential and non-referential values. However, since this is an analysis of isolated entities, there are no non-coreferential links. For that reason, in this paper we only present *coreferential* precision, recall and F-score for this metric: $P_c = \frac{rc}{rc+wc}$, $R_c = \frac{rc}{rc+wn}$ and $F_c = \frac{2P_cR_c}{P_c+R_c}$, where *rc* is the number of correct coreferential links, *wc* the number of incorrect coreferential links, and *wn* is the number of non-coreferential links incorrectly marked as coreferent by the system.

2.2 Entity and mention-based: B^3 and CEAF

B^3 (Bagga and Baldwin, 1998) calculates precision and recall for every mention in the document, and then combines them to an overall precision and recall. Precision of a single mention m_i is the number of correct mentions in the response entity R_i that contains m_i divided by the total number of mentions in R_i. Recall of m_i is again the number of correct mentions in R_i, this time divided by the number of mentions in the key entity K_i that contains mention m_i. The precision and recall for the entire document can be calculated as weighted sums of precision and recall of the individual mentions. The default weight, also used in this experiment is $\frac{1}{n}$, where n is the number of mentions in the document.

CEAF (Luo, 2005) is based on the best alignment of subsets of key and response entities. For any mapping $g \in G_m$ the total similarity $\Phi(g)$ is the sum of all similarities. The best alignment g^* is found by maximizing the sum of similarities $\Phi(g)$ between the key and response entities, while the maximum total similarity is the sum of the best similarities. Precision and recall are defined in terms of the similarity measure $\Phi(g^*)$: $P = \frac{\Phi(g^*)}{\sum_i \phi(R_i,R_i)}$ $R = \frac{\Phi(g^*)}{\sum_i \phi(K_i,K_i)}$.

There are two versions of CEAF with different similarity measures, $\phi_m(K,R) = |K \cap R|$ and $\phi_e(K,R) = \frac{2|K \cap R|}{|K|+|R|}$. ϕ_e is the basis for CEAF_e which shows a measure of correct entities while CEAF_m, based on ϕ_m, shows the percentage of correctly resolved mentions.

	MUC	B^3	CEAF_e	CEAF_m	BLANC	MELA
MUC	–	0.46	0.22	0.47	0.35	0.63
B^3	0.59	–	0.47	0.56	0.42	0.61
CEAF_e	0.46	0.59	–	0.51	0.26	0.38
CEAF_m	0.57	0.70	0.62	–	0.46	0.60
BLANC	0.57	0.70	0.57	0.68	–	0.35
MELA	0.59	0.73	0.59	0.70	0.70	–

Table 1: Kendall τ rank correlation coefficient for teams participating in CoNLL shared tasks, with CoNLL 2011 in the upper right, CoNLL 2012 in the lower left corner.

3 Correlating CoNLL shared tasks results

To illustrate the complexity of the present evaluation best practices, we have applied the Kendall τ rank correlation coefficient to the ratings the metrics gave coreference resolution systems that competed in the two recent CoNLL shared tasks. The official metrics of the CoNLL shared tasks was MELA (Denis and Baldridge, 2009), a weighted average of MUC, B^3 and CEAF_e.

The results for CoNLL 2011 (Table 1) show a rather weak correlation among the metrics going down to as low as 0.22 between CEAF_e and MUC. Somewhat surprisingly, the two link-based metrics, MUC and BLANC, also show a low degree of agreement (0.35), while the mention-based metrics, CEAF_m and B^3, show the highest agreement of all non-composite metrics. However, this agreement is not particularly high either as the two metrics agree on just above the half of all the cases (0.56).

The results for CoNLL 2012 show much higher correlation among the metrics ranging from 0.46 to 0.70. Again CEAF_m and B^3 show the highest correlation, but unlike in 2011 BLANC "joins" this cluster. CEAF_e and MUC are again least correlated, while CEAF_e and BLANC, in 2011 almost independent, show average correlation (0.57) in 2012.

In our view, comparatively low correlations as well as surprising variation from year to year suggests a certain degree of 'fuzziness' in quantitative coreference resolution evaluation. We leave the investigation of variation between the two years for future work.

4 Error analysis

To better understand the functioning of the metrics we have conducted an error analysis on the key/response entity pairs from five short texts from

the development corpus of the CoNLL 2011 Shared Task (Pradhan et al., 2011), one text from each of the five represented genres: Broadcast Conversations (BC), Broadcast News (BN), Magazine (MZ), News Wire (NW) and Web Blogs and News Groups (WB). The texts were chosen as randomly as possible, the only constraint being length[1]. The gold standard texts are originally from OntoNotes 4.0, and contain 64 mentions distributed among 21 key entities. The response texts are the output of Stanford's Multi-Pass Sieve Coreference Resolution System (Lee et al., 2011).

4.1 Categorization

Instead of classifying entities according to their score by some of the metrics, or a combination of several of them, as done by the CoNLL shared tasks, we have based the classification on a notion of linguistic common sense – our subjective idea of how humans evaluate the success or failure of CR. We divide key/response entity pairs into four categories:

- Category 1: Perfect match
- Category 2: Partial match
- Category 3: Merged entities
- Category 4: Failed coreference resolution

We will concentrate on the amount of informational value from the key entity that has been preserved in the response entity. In the course of these experiments, our aim is to see if that rather informal idea can be operationalized in a way amenable to future use in automated CR and/or quantitative evaluation.

4.1.1 Category 1: Perfect match

This class consists of four key/response entity pairs with complete string match. The key and response entities being identical, all metrics show unanimously precision and recall of 100%. The informational value is, of course, completely preserved. Unfortunately, those examples are few and simple: They constitute only 19% of the entities and 14% of mentions in this sample, and all seem to be achieved by the simplest form of string matching.

Key entities	Response entities		MUC	B^3	$CEAF_e$	$CEAF_m$	BLANC
BC45							
• The KMT vice chairman • Wang Jin-pyng • his • his • He • he	• Wang Jin-pyng • his • his • He • he	P R F	100.00 80.00 88.89	100.00 83.33 90.91	90.90 90.90 90.90	100.00 83.33 90.91	100.00 66.67 80.00
BC22							
• KMT Chairman Lien Chan • Chairman Lien Chan • Lien Chan	• KMT Chairman Lien Chan • Lien Chan	P R F	100.00 50.00 66.67	100.00 66.67 80.00	80.00 80.00 80.00	100.00 66.67 80.00	100.00 33.33 50.00
BN1							
• Bill Clinton • The President • he • his • Mr.Clinton • his • He • he	• Bill Clinton • Mr.Clinton • his • He • he	P R F	100.00 85.71 92.31	100.00 62.50 76.92	76.92 76.92 76.92	100.00 62.50 76.92	68.42 46.43 55.32
NW2							
• New Zealand • New Zealand • New Zealand • New Zealand's • New • Zealand	• New Zealand • New Zealand • New Zealand • New Zealand's	P R F	100.00 75.00 85.71	100.00 80.00 88.89	88.89 88.89 88.89	100.00 80.00 88.89	100.00 60.00 75.00

Table 2: Category 2a: Partial match (partial entities)

4.1.2 Category 2: Partial match

The partial response entities can be divided in two subcategories: **2a)** The cases where the response *entities* are partial, i.e.they form a proper subset of the key entity mentions (Table 2) and **2b)** The cases where the response *mentions* are partial, i.e. substrings of the corresponding key mentions (Table 3).

The scoring of the examples has followed CoNLL shared tasks' strict mention detection requirements[2] with the consequence that Category 2b entities have received considerably lower scores than the Category 2a entities even in cases where the loss of informational value has been comparable. For instance, the response entity NW1 (Table 3) has received an average F-score of 56.67%, but its loss of informational value is comparable to that in entities BC45 and BN1 (Table 2). The BC45's response entity has lost the information that `Jiyun Tian` is a `vice-chief`, while entities BC45 and BN1 have lost the information that the person referred to

[1]The texts longer than five sentences were discarded, to make the analysis tractable.

[2]Only response mentions with boundaries identical to the gold mentions are recognized as correct (Pradhan et al., 2011)

is `The KMT vice chairman` (BC45) and `The President` (BN1). However, the latter mentions have received a considerably higher average F-score of 88.32% and 75.68% respectively. This indicates that stricter mention detection requirements do not necessarily improve the quality of CR evaluation.

Key entities	Response entities		MUC	B^3	$CEAF_e$	$CEAF_m$	BLANC
MZ22							
• a school in Shenzhen for the children of Hong Kong expats • the school in Shenzhen	• a school in Shenzhen • the school in Shenzhen	P R F	0.00 0.00 0.00	50.00 50.00 50.00	50.00 50.00 50.00	50.00 50.00 50.00	0.00 0.00 0.00
NW0							
• China's People's Congress • China's People's Congress	• People's Congress • People's Congress	P R F	0.00 0.00 0.00	0.00 0.00 0.00	0.00 0.00 0.00	0.00 0.00 0.00	0.00 0.00 0.00
NW1							
• vice-chief committee member Jiyun Tian • Jiyun Tian • He	• committee member Jiyun Tian • Jiyun Tian • He	P R F	50.00 50.00 50.00	66.67 66.67 66.67	66.67 66.67 66.67	66.67 66.67 66.67	33.33 33.33 33.33
NW5							
• China's People's Congress delegation led by vice-chief committee member Jiyun Tian • the delegation from China's People's Congress	• China's People's Congress delegation • the delegation from China's People's Congress	P R F	0.00 0.00 0.00	50.00 50.00 50.00	50.00 50.00 50.00	50.00 50.00 50.00	0.00 0.00 0.00

Table 3: Category 2b: Partial match (partial mentions).

4.1.3 Category 3: Merged entities

This category consists of response entities that contain mentions from two or more key entities (Table 4). Our sample contains only four examples in this category, but it is still possible to discern two subcategories:

1. The new information is incorrect

In the key entity MZ40, the sex of the gender-neutral `her ten-year-old child` has been given by the mention `him`. Replacing it with the mention `she` in the response entity gives the wrong information about the child's sex. Entities BN2 and MZ17 also belong to this subcategory, but here the mentions in the response entity are morphologically inconsistent, thus making the mistake easier to detect.

2. The new information is correct or neutral

In entity pair MZ19 the key mention `the latter group` was replaced with response mention `them`,

Key entities	Response entities		MUC	B^3	$CEAF_e$	$CEAF_m$	blanc
BN2							
• he and his wife, now a New York senator • their	• The President • he and his wife, now a New York senator • he • his	P R F	66.67 0.00 0.00	25.00 50.00 33.33	33.33 33.33 33.33	25.00 50.00 33.33	0.00 0.00 0.00
MZ19							
• the more affluent Taiwanese • their • the latter group	• the more affluent Taiwanese • their • them	P R F	50.00 50.00 50.00	66.67 66.67 66.67	66.67 66.67 66.67	66.67 66.67 66.67	33.33 33.33 33.33
MZ17							
• her elder son and daughter • them	• her elder son and daughter • him • him	P R F	0.00 0.00 0.00	33.33 50.00 40.00	40.00 40.00 40.00	33.33 50.00 40.00	0.00 0.00 0.00
MZ40							
• Her ten-year-old child • him • The child • him	• Her ten-year-old child • she • The child	P R F	50.00 33.33 40.00	66.67 50.00 57.14	57.14 57.14 57.14	66.67 50.00 57.14	33.33 20.00 25.00

Table 4: Category 3: Merged entities

the omitted and replacement mentions having very similar informational content.

As expected, the scores in Category 3 are lower then those in Category 2 (as a whole), but they are still consistently better than the scores of the Category 2b.

4.1.4 Category 4: Unsuccessful coreference resolution

The entities in this category (Table 5) are divided into two subcategories:

No response entity has been given Two of the key entities (MZ38 and NW4) were not aligned with any response entities, and not surprisingly all metrics agree that the CR precision, recall and F-score equal zero.

The response entities do not contain a single "heavy" mention that is correct Although the response entities in the remaining entity pairs are non-empty, an intuitive CR evaluation says there is not much sense in aligning near-vacuous mentions if the entity is otherwise wrong or empty. Already in the two rather simple cases of WB0 and WB1 the metrics show large discrepancies: While link-based MUC and BLANC correctly give an F-score of 0.00 as there are no correct links in the entity, the mention-based B^3 and CEAF measures award them

Key entities	Response entities		MUC	B^3	$CEAF_e$	$CEAF_m$	BLANC
WB0							
• the beauty industry • it	• the one hand • it	P	0.00	50.00	50.00	50.00	0.00
		R	0.00	50.00	50.00	50.00	0.00
		F	0.00	50.00	50.00	50.00	0.00
WB1							
• the consumer • they	• clinical dermatologists • they	P	0.00	50.00	50.00	50.00	0.00
		R	0.00	50.00	50.00	50.00	0.00
		F	0.00	50.00	50.00	50.00	0.00
MZ33							
• Chang, Mei-liang, chairperson of the TBAD Women's Division, • her • she • Her • she	• her • she • Her	P	100.00	100.00	75.00	100.00	100.00
		R	50.00	60.00	75.00	60.00	30.00
		F	66.67	75.00	75.00	75.00	46.15

Table 5: Category 4: Unsuccessful coreference resolution

Entity	MUC	B^3	$CEAF_e$	$CEAF_m$	BLANC	Human
BC45	6	5	5	5	5	7
BC22	8	7	7	7	8	8
BC51	1	1	1	1	1	1
BN0	1	1	1	1	1	1
BN1	5	8	8	8	7	13
BN2	**13**	18	18	18	**13**	15
MZ19	10	10	10	10	10	14
MZ33	8	9	9	9	9	17
MZ17	**13**	17	17	17	**13**	15
MZ40	12	12	12	12	12	17
MZ22	**13**	13	13	13	**13**	10
MZ24	1	1	1	1	1	1
MZ38	**–**	**19**	**19**	**19**	**13**	**19**
NW0	**13**	**19**	**19**	**19**	**13**	10
NW1	10	10	10	10	10	8
NW2	7	6	6	6	6	5
NW3	1	1	1	1	1	1
NW4	**–**	**19**	**19**	**19**	**13**	**19**
NW5	**13**	13	13	13	**13**	10
WB0	**13**	13	13	13	**13**	**19**
WB1	**13**	13	13	13	**13**	**19**

Table 6: Ranking of our example entities.

with a rather high F-score of 50.00.

Entity MZ33 has been awarded high F-scores by all metrics, averaging 67.56%. However, almost all information from the key entity in MZ33 has been lost in the response entity: The key entity contains information on a person, a female, a Taiwanese national, her name (`Chang Mei-lian`) and the additional information that she is a `chairperson of the TBAD, Women's Division`. The response entity contains the information that its mentions refer to a female, which is most probably a person, but might be a ship, or a well loved pet. None of the metrics indicate that such a substantial loss of information renders the coreference resolution of MZ33 practically useless for a human user.

5 Entity ranking

As some of the metrics yield consistently lower F-score levels, it is more appropriate to compare rankings of entities than the actual F-scores (Table 6). We have also – to infuse an iota of old-school armchair linguistics – added a sixth rating column, showing intuitive rankings, based on informational value retained. The lowest rankings for any metric are marked in **bold**.

The entities showing broad agreement among the metrics are only the best (Category 1) and the worst ones (MZ38 and NW4, Category 4).

The metrics disagreement surfaces with entities WB0 and WB1 of Category 4. The link-based metrics, MUC and BLANC, rank them last (13^{th}), while they are ranked much higher (13^{th} out of 19) by the mention-based and entity-based metrics (B^3 and CEAF). In this case the human evaluator agrees with the link-based metrics: If there is not a single correct link within an entity, our intuition says that no useful CR has taken place.

However, the presence of a single correct coreferent link is not sufficient for our intuition of successful resolution. Consider entities MZ22 and NW5 (Table 3): They also consist of two entities where only one is correct, and have received the same ratings as WB0 and WB1, but in this case we judge CR as much more successful. There are two main differences between this and the previous case. Firstly, the correct mention is in the previous case a meaning-lean pronoun (`it` and `they`) while the correct mention in this case is a 'full-bodied' NP (`the school in Shenzhen` and `the delegation from China's People's Congress`). In addition, in both of the Category 2 entity pairs, the incorrect mention holds an informational value very close or identical to that of the correct mention. This example illustrates the importance of informational value content of the mentions for the human evaluation of the resolution.

6 Formalizing the intuition

We have earlier (§4.1) introduced a classificaion based on an informal notion of (human) intuitive coreference resolution evaluation. In this section we will try to formalize the classification.

Category 1 The key entities and response entities are identical:

$$\forall x(x \in K \leftrightarrow x \in R) \tag{1}$$

Category 2 The response entity is a proper subset of the key entity:

$$\begin{aligned} &\forall x(x \in R \rightarrow x \in K) \wedge \\ &\exists y(y \in K \wedge y \notin R) \end{aligned} \tag{2a}$$

This is the only condition for Category 2a. Category 2b shares the condition (2a), but to formalize it, we have to add *overlap(x,y)* relation. We can define it as a common substring for x and y of a certain length, possibly including at least one major syntactic category, or even the lexical 'head' if some way of operationalizing that notion is available.

$$\begin{aligned} &\exists x(x \in K \wedge x \in R) \wedge \\ &\exists y \exists z(y \in K \wedge z \in R \wedge overlap(y, z)) \end{aligned} \tag{2b}$$

We need at least two correct mentions in the response entity, and at least one that overlaps, as response entities containing only one correct mention do not have any correct links.

Category 3 Response entity contains a subset of the key entity mentions as well as additional mention(s) belonging to some other entity (E):

$$\begin{aligned} &\exists x(x \in K \wedge x \in R) \wedge \\ &\exists y(y \notin K \wedge y \in R \wedge y \in E) \end{aligned} \tag{3}$$

Category 4 The entities belonging to this category have a twofold definition: The response entity is either empty or if it contains one correct mention, it cannot contain an overlapping mention.

$$\begin{aligned} &\forall x(x \in K \rightarrow x \notin R) \vee \\ &\exists x(x \in K \wedge x \in R) \rightarrow \\ &\forall y \forall z((y \in K \wedge z \in R) \rightarrow \neg overlap(y, z)) \end{aligned} \tag{4}$$

The classification that has been introduced as a informal one in §4.1 is thus computable given an operational definition of overlap. In future work we will investigate the distribution of the four error categories on a larger sample.

7 Conclusion and outlook

In this paper we have compared metrics on the basis of their evaluation of coreference resolution performed on real-life texts, and contrasted their evaluation to an intuitive human evaluation of coreference resolution. We conjecture that humans require both correct coreferent links and correct (whole or partial) mentions of a certain information weight to consider a resolution successful.

This approach has some shortcomings. Firstly, the manual nature of the analysis has imposed a limit on the number of the examples, so our data may not be representative. Secondly, there is uncertainity connected to how well the coreference resolution evaluation metrics are suited to be used in this way. The latter drawback is the more serious one: the metrics were not designed to evaluate single key/response pairs, but whole texts. However, we would argue that if we want to discover new insights into the evaluation process, some level of approximation is necessary. There are at least two arguments in favor of this particular approximation: Firstly, all metrics are based on evaluating key/response pairs. Analyzing their performance at this level can be a reasonable indicator of their performance on the text level. Secondly, even if metrics are treated "unfairly", they are all treated equally.

We thus believe that this work can be seen as an illustration of remaining evaluation challenges in the field of coreference resolution.

A natural extension of this work would be including more humans in evaluating coreference resolution systems, to provide a more representative human judgement. This evaluation should then be extended from evaluating coreference resolution of single key/response entity pairs, to assessing the quality of coreference resolution on a text as a whole.

And, finally: Every mention carries an information value, and this weight varies from quite heavy (as in `vice-chief committee member Jiyun Tian`), to somewhat lighter (`Jiyun Tian`) to virtually weightless (`He`). Information weights are not distributed randomly, but conform to discourse structure. It would be interesting to map the pattern of their distribution, and see if incorporating this information could improve both coreference resolution and its quantitative evaluation.

8 Acknowledgements

We would like to thank the anonymous reviewers for their useful comments.

References

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, pages 563 – 566, Granada, Spain.

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

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2011. Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 Shared Task. In *Proceedings of the 15th Conference on Computational Natural Language Learning: Shared Task*, pages 28–34, Portland, Oregon.

Xiaoqiang Luo, Abe Ittycheriah Hongyan Jing, Nanda Kambhatla, and Salim Roukos. 2004. A mention-synchronous coreference resolution algorithm based on the Bell tree. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 136–143, Barcelona, Spain.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, pages 25–32, Vancouver, Canada.

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 15th Conference on Computational Natural Language Learning: Shared Task*, pages 1–27, Portland, Oregon.

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 the Joint Conference on EMNLP and CoNLL: Shared Task*, pages 1–40, Jeju Island, Korea.

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

Marta Recasens and Eduard Hovy. 2011. BLANC: Implementing the Rand index for coreference evaluation. *Natural Language Engineering*, 17(04):485–510.

Marta Recasens, Lluis Màrquez, Emili Sapena, M. Antònia Marti, Mariona Taulé, Véronique Hoste, Massimo Poesio, and Yannick Versley. 2010. Semeval-2010 task 1: Coreference resolution in multiple languages. In *Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010*, pages 1–8, Uppsala, Sweden.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of the Sixth Message understanding Conference (MUC-6)*, pages 45–52, San Francisco, CA.

Xiaofeng Yang, Guodong Zhou, Jian Su, and Chew Lim Tan. 2003. Coreference resolution using competition learning approach. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 176–183, Sapporo, Japan.