CONE: Metrics for Automatic Evaluation of Named Entity Co-reference Resolution

Bo Lin, Rushin Shah, Robert Frederking, Anatole Gershman
Language Technologies Institute, School of Computer Science
Carnegie Mellon University
5000 Forbes Ave., PA 15213, USA
{bolin,rnshah,ref,anatoleg}@cs.cmu.edu

Abstract

Human annotation for Co-reference Resolution (CRR) is labor intensive and costly, and only a handful of annotated corpora are currently available. However, corpora with Named Entity (NE) annotations are widely available. Also, unlike current CRR systems, state-of-the-art NER systems have very high accuracy and can generate NE labels that are very close to the gold standard for unlabeled corpora. We propose a new set of metrics collectively called CONE for Named Entity Co-reference Resolution (NE-CRR) that use a subset of gold standard annotations, with the advantage that this subset can be easily approximated using NE labels when gold standard CRR annotations are absent. We define CONE B^3 and CONE CEAF metrics based on the traditional B^3 and CEAF metrics and show that CONE B^3 and CONE CEAF scores of any CRR system on any dataset are highly correlated with its B^3 and CEAF scores respectively. We obtain correlation factors greater than 0.6 for all CRR systems across all datasets, and a best-case correlation factor of 0.8. We also present a baseline method to estimate the gold standard required by CONE metrics, and show that CONE B^3 and CONE CEAF scores using this estimated gold standard are also correlated with B^3 and CEAF scores respectively. We thus demonstrate the suitability of CONE $B^3$and CONE CEAF for automatic evaluation of NE-CRR.

1 Introduction

Co-reference resolution (CRR) is the problem of determining whether two entity mentions in a text refer to the same entity in real world or not. Noun Phrase CRR (NP-CRR) considers all noun phrases as entities, while Named Entity CRR restricts itself to noun phrases that describe a Named Entity. In this paper, we consider the task of Named Entity CRR (NE-CRR) only. Most, if not all, recent efforts in the field of CRR have concentrated on machine-learning based approaches. Many of them formulate the problem as a pair-wise binary classification task, in which possible co-reference between every pair of mentions is considered, and produce chains of co-referring mentions for each entity as their output. One of the most important problems in CRR is the evaluation of CRR results. Different evaluation metrics have been proposed for this task. B-cubed (Bagga and Baldwin, 1998) and CEAF (Luo, 2005) are the two most popular metrics; they compute Precision, Recall and F1 measure between matched equivalent classes and use weighted sums of Precision, Recall and F1 to produce a global score. Like all metrics, B^3 and CEAF require gold standard annotations; however, gold standard CRR annotations are scarce, because producing such annotations involves a substantial amount of human effort since it requires an in-depth knowledge of linguistics and a high level of understanding of the particular text. Consequently, very few corpora with gold standard CRR annotations are available (NIST, 2003; MUC-6, 1995; Agirre, 2007). By contrast, gold standard Named Entity (NE) annotations are easy to produce; indeed, there are many NE annotated corpora of different sizes and genres. Similarly, there are few CRR systems and even the best scores obtained by them are only in the region of F1 = 0.5 - 0.6. There are only four such CRR systems freely available, to the best of our knowledge (Bengston and Roth, 2007; Versley et al., 2008; Baldridge and Torton, 2004; Baldwin and Carpenter, 2003). In comparison, there are numerous Named Entity recognition (NER) systems, both general-purpose and specialized, and many of them achieve scores better than F1 = 0.95 (Ratinov and Roth, 2009; Finkel et al.,

Proceedings of the 2010 Named Entities Workshop, ACL 2010, pages 136–144,
Uppsala, Sweden, 16 July 2010. ©2010 Association for Computational Linguistics

2005). Although these facts can be partly attributed to the 'hardness' of CRR compared to NER, they also reflect the substantial gap between NER and CRR research. In this paper, we present a set of metrics, collectively called CONE, that leverage widely available NER systems and resources and tools for the task of evaluating co-reference resolution systems. The basic idea behind CONE is to predict a CRR system's performance for the task of full NE-CRR on some dataset using its performance for the subtask of named mentions extraction and grouping (NMEG) on that dataset. The advantage of doing so is that measuring NE-CRR performance requires the co-reference information of all mentions of a Named Entity, including named mentions, nominal and pronominal references, while measuring the NMEG performance only requires co-reference information of named mentions of a NE, and this information is relatively easy to obtain automatically even in the absence of gold standard annotations. We compute correlation between CONE B^3, B^3, CONE CEAF and CEAF scores for various CRR systems on various gold-standard annotated datasets and show that the CONE B^3 and B^3 scores are highly correlated for all such combinations of CRR systems and datasets, as are CONE CEAF and CEAF scores, with a best-case correlation of 0.8. We produce estimated gold standard annotations for the Enron email corpus, since no actual gold standard CRR annotations exist for it, and then use CONE B^3 and CONE CEAF with these estimated gold standard annotations to compare the performance of various NE-CRR systems on this corpus. No such comparison has been previously performed for the Enron corpus.

We adopt the same terminology as in (Luo, 2005): a *mention* refers to each individual phrase and an *entity* refers to the *equivalence class* or *co-reference chain* with several mentions. This allows us to note some differences between NE-CRR and NP-CRR. NE-CRR involves indentifying named entities and extracting their co-referring mentions; equivalences classes without any NEs are not considered. NE-CRR is thus clearly a subset of NP-CRR, where all co-referring mentions and equivalence classes are considered. However, we focus on NE-CRR because it is currently a more active research area than NP-CRR and a better fit for target applications such as text forensics and web mining, and also because it is more amenable to the automatic evaluation approach that we propose.

The research questions that motivate our work are:

(1) Is it possible to use only NER resources to evaluate NE-CRR systems? If so, how is this problem formulated?
(2) How does one perform evaluation in a way that is accurate and automatic with least human intervention?
(3) How does one perform evaluation on large unlabeled datasets?

We show that our CONE metrics achieve good results and represent a promising first step toward answering these questions.

The rest of the paper is organized as follows. We present related work in the field of automatic evaluation methods for natural language processing tasks in Section 2. In Section 3, we give an overview of the standard metrics currently used for evaluating co-reference resolution. We define our new metrics CONE B^3 and CONE CEAF in Section 4. In section 5, we provide experimental results that illustrate the performance of CONE B^3 and CONE CEAF compared to B^3 and CEAF respectively. In Section 6, we give an example of the application of CONE metrics by evaluating NE-CRR systems on an unlabeled dataset, and discuss possible drawbacks and extensions of these metrics. Finally, in section 7 we present our conclusions and ideas for future work.

2 Related Work

There has been a substantial amount of research devoted to automatic evaluation for natural language processing, especially tasks involving language generation. The BLEU score (Papineni et al., 2002) proposed for evaluating machine translation results is the best known example of this. It uses n-gram statistics between machine generated results and references. It inspired the ROUGE metric (Lin and Hovy, 2003) and other methods (Louis and Nenkova, 2009) to perform automatic evaluation of text summarization. Both these metrics have show strong correlation between automatic evaluation results and human judgments. The two metrics successfully reduce the need for human judgment and help speed up research by allowing large-scale evaluation. Another example is the alignment entropy (Pervouchine et al., 2009) for evaluating transliteration alignment. It reduces the need for alignment gold standard and highly correlates with transliteration system performance. Thus it is able to serve as a good metric for transliteration alignment. We contrast our work with (Stoyanov et al., 2009), who show that the co-reference resolution problem can be separated into different parts according to the type of the mention. Some parts are relatively easy to solve. The resolver performs equally well in each part across datasets. They use the statistics of mentions in different parts with test results on other datasets as a predictor for unseen datasets, and obtain promising results with good correlations. We approach the problem from a different perspective. In our work, we show the correlation between the scores on traditional metrics and scores on our CONE metrics, and show how to automatically estimate the gold standard required by CONE metrics. Thus our method is able to predict the co-reference resolution performance without gold standard at all. We base our new metrics on the standard B^3 and CEAF metrics used for computing CRR scores. (Vilian et al., 1995; Bagga and Baldwin, 1998; Luo, 2005). B^3 and CEAF are believed to be more discriminative and interpretable than earlier metrics and are widely adopted especially for machine-learning based approaches.

3 Standard Metrics: B^3 and CEAF

We now provide an overview of the standard B^3 and CEAF metrics used to evaluate CRR systems. Both metrics assume that a CRR system produces a set of equivalence classes {*O*} and assigns each mention to only one class. Let O_i be the class to which the *i*th mention was assigned by the system. We also assume that we have a set of correct equivalence classes {*G*} (the gold standard). Let G_i be the gold standard class to which the *i*th mention should belong. Let N_i denote the number of mentions in O_i which are also in G_i – the correct mentions. B^3 computes the presence rate of correct mentions in the same equivalent classes. The individual precision and recall score is defined as follows:

$$P_i = \frac{N_i}{|O_i|} \qquad R_i = \frac{N_i}{|G_i|}$$

Here $|O_i|$ and $|G_i|$ are the cardinalities of sets O_i and G_i.

The final precision and recall scores are:

$$P = \sum_{i=1}^{n} w_i P_i \qquad R = \sum_{i=1}^{n} w_i R_i$$

Here, in the simplest case the weight w_i is set to 1/n, equal for all mentions.

CEAF (Luo, 2005) produces the optimal matching between output classes and true classes first, with the constraint that one true class, G_i, can be mapped to at most one output class, say $O_{f(i)}$ and vice versa. This can be solved by the KM algorithm (Kuhn, 1955; Munkres, 1957) for maximum matching in a bipartite graph. CEAF then computes the precision and recall score as follows:

$$P = \frac{\sum_i |M_{i,f(i)}|}{\sum_i |O_i|} \qquad R = \frac{\sum_i |M_{i,f(i)}|}{\sum_i |G_i|}$$

$$M_{i,j} = |O_i \cap G_j|$$

We use the terms $M_{i,j}$ from CEAF to re-write B^3, its formulas then reduce to:

$$P = \frac{1}{\sum_i |O_i|} \sum_i \sum_j \frac{|M_{i,j}|^2}{|O_i|}$$

$$R = \frac{1}{\sum_i |G_i|} \sum_i \sum_j \frac{|M_{i,j}|^2}{|G_i|}$$

We can see that B^3 simply iterates through all pairs of matchings instead of considering the one to one mappings as CEAF does. Thus, B^3 computes the weighted sum of the F-measures for each individual mention which helps alleviate the bias in the pure link-based F-measure, while CEAF computes the same as B^3 but enforces at most one matched equivalence class for every class in the system output and gold standard output.

4 CONE B^3 and CONE CEAF Metrics:

We now formally define the new CONE B^3 and CONE CEAF metrics that we propose for automatic evaluation of NE-CRR systems.

Let *G* denote the set of gold standard annotations and *O* denote the output of an NE-CRR system. Let G_i denote the equivalent class of entity *i* in the gold standard and O_j denote the equivalence class for entity *j* in the system output. Also let G_{ij} denote the jth mention in the equivalence class of entity i in the gold standard and O_{ij} denote the jth mention in the system output.

As described earlier, the standard B^3 and CEAF metrics evaluate scores using *G* and *O* and can be thought of as functions of the form $B^3(G, O)$. and *CEAF*(*G*, *O*) respectively. Let us use *Score*(*G*, *O*) to collectively refer to both these functions. An equivalence class G_i in G may contain three types of mentions: named mentions $g^{NM}{}_{ij}$, nominal mentions $g^{NO}{}_{ij}$, and pronominal mentions $g^{PR}{}_{ij}$. Similarly, we can define $o^{NM}{}_{ij}$, $o^{NO}{}_{ij}$ and $o^{PR}{}_{ij}$ for a class O_i in O. Now for each gold standard equivalence class G_i and system output equivalence class O_i, we define the following sets $G^{NM}{}_i$ and $O^{NM}{}_i$:

$$\forall i, G^{NM}{}_i = \{g^{NM}{}_{ij}\}, g^{NM}{}_{ij} \subset G_i$$

$$\forall i, O^{NM}{}_i = \{o^{NM}{}_{ij}\}, o^{NM}{}_{ij} \subset O_i$$

In other words, $G^{NM}{}_i$ and $O^{NM}{}_i$ are the subsets of G_i and O_i containing all named mentions and no mentions of any other type.

Let G^{NM} denote the set of all such equivalance classes $G^{NM}{}_i$ and O^{NM} denote the set of all equivalence classes $O^{NM}{}_i$. It is clear that G^{NM} and O^{NM} are pruned versions of the gold standard annotations and system output respectively.

We now define CONE B^3 and CONE CEAF as follows:

CONE B^3 = B^3(G^{NM}, O^{NM})
CONE CEAF = CEAF(G^{NM}, O^{NM})

Following our previous notation, we denote CONE B^3 and CONE CEAF collectively as Score(G^{NM}, O^{NM}). We observe that Score(G^{NM}, O^{NM}) measures a NE-CRR system's performance for the NE-CRR subtask of named mentions extraction and grouping (NMEG). We find that Score(G^{NM}, O^{NM}) is highly correlated with Score(G, O) for all the freely available NE-CRR systems over various datasets. This provides the neccessary justification for the use of Score(G^{NM}, O^{NM}).

We use SYNERGY (Shah et al., 2010), an ensemble NER system that combines the UIUC NER (Ritanov and Roth, 2009) and Stanford NER (Finkel et al., 2005) systems, to produce G^{NM} and O^{NM} from G and O by selecting named mentions. However, any other good NER system would serve the same purpose.

We see that while standard evaluation metrics require the use of G, i.e. the full set of NE-CRR gold standard annotations including named, nominal and pronimal mentions, CONE metrics require only G^{NM}, i.e. gold standard annotations consisting of named mentions only. The key advantage of using CONE metrics is that G^{NM} can be automatically approximated using an NER system with a good degree of accuracy. This is because state-of-the-art NER systems achieve near-optimal performance, exceeding F1 = 0.95 in many cases, and after obtaining their output, the task of estimating G^{NM} reduces to simply clustering it to seperate mentions of diffrerent real-world entities. This clustering can be thought of as a form of named entity matching, which is not a very hard problem. There exist systems that perform such matching in a sophisticated manner with a high degree of accuracy. We use simple heuristics such as exact matching, word matches, matches between initials, etc. to design such a matching system ourselves and use it to obtain estimates of G^{NM}, say $G^{NM\text{-}approx}$. We then calculate CONE B^3 and CONE CEAF scores using $G^{NM\text{-}approx}$ instead of G^{NM}; in other words, we perform fully automatic evaluation of NE-CRR systems by using Score($G^{NM\text{-}approx}$, O^{NM}) instead of Score(G^{NM}, O^{NM}). In order to show the validity of this evaluation, we calculate the correlation between the Score($G^{NM\text{-}approx}$, O^{NM}) and Score(G, O) for different NE-CRR systems across different datasets and find that they are indeed correlated. CONE thus makes automatic evaluation of NE-CRR systems possible. By leveraging the widely available named entity resources, it reduces the need for gold standard annotations in the evaluation process.

4.1 Analysis

There are two major kinds of errors that affect the performance of NE-CRR systems for the full NE-CRR task:

- Missing Named Entity (MNE): If a named mention is missing from the system output, it is very likely that its nearby nominal and anaphoric mentions will be lost, too
- Incorrectly grouped Named Entity (IGNE): Even if the named mention is correctly identified with its nearby nominal and anaphoric mentions to form a chain, it is still possible to misclassify the named mentions and its co-reference chain

Consider the following example of these two types of errors. Here, the alphabets represent the named mentions and numbers represent other type of mentions:

Gold standard, *G*: (A, B, C, 1, 2, 3, 4)
Output from System 1, *O1*: (A, B, 1, 2, 3)
Output from System 2, *O2*: (A, C, 1, 2, 4), (B, 3)
O1 shows an example of an MNE error, while *O2* shows an example of an IGNE error.

Both these types of errors are in fact rooted in named mention extraction and grouping (NMEG). Therefore, we hypothesize that they must be preserved in a NE-CRR system's output for the subtask of named mentions extraction and grouping (NMEG) and will be reflected in the CONE B^3 and CONE CEAF metrics that evaluate scores for this subtask. Consider the following extension of the previous example:

G^{NM}: (A, B, C)
$O1^{NM}$: (A, B)
$O2^{NM}$: (A, C), (B)

We observe that the MNE error in *O1* is preserved in $O1^{NM}$, and the IGNE error in *O2* is preserved in $O2^{NM}$. Empirically we sample several output files in our experiments and observe the same phenomena. Therefore, we argue that it is possible to capture the two major kinds of errors described by considering only G^{NM} and O^{NM} instead of *G* and *O*.

We now provide a more detailed theoretical analysis of the CONE metrics. For a given NE-CRR system and dataset, consider the system output *O* and gold standard annotation *G*. Let P and R indicate precision and recall scores obtained by evaluating *O* against *G*, using CEAF. If we replace both *G* and *O* with their subsets G^{NM} and O^{NM} respectively, such that G^{NM} and O^{NM} contain only named mentions, we can modify the equations for precision and recall for CEAF to derive the following equations for precision P^{NM} and recall R^{NM} for CONE CEAF:

$$Sum\{O^{NM}\} = \sum_i \left|O^{NM}{}_i\right|$$

$$Sum\{G^{NM}\} = \sum_i \left|G^{NM}{}_i\right|$$

$$P^{NM} = \sum_i \frac{\left|M^{NM}{}_{i,f(i)}\right|}{Sum\{O^{NM}\}}$$

$$R^{NM} = \sum_i \frac{\left|M^{NM}{}_{i,f(i)}\right|}{Sum\{G^{NM}\}}$$

The corresponding equations for CONE B^3 Precision are:

$$P^{NM} = \sum_i \frac{\sum_j \left|M^{NM}{}_{i,j}\right|^2}{\left|O^{NM}{}_i\right| \times Sum\{O^{NM}\}}$$

$$R' = \sum_i \frac{\sum_j \left|M^{NM}{}_{i,j}\right|^2}{\left|R^{NM}{}_i\right| \times Sum\{R^{NM}\}}$$

In order to support the hypothesis that CONE metrics evaluated using (G^{NM}, O^{NM}) represent an effective substitute for standard metrics that use (*G*, *O*), we compute entity level correlation between the corresponding CONE and standard metrics. For example, in the case of CEAF / CONE CEAF Precision, we calculate correlation between the following quantities:

$$\vec{P}^{NM} = < \frac{\left|M^{NM}{}_{i,f(i)}\right|}{Sum\{S^{NM}\}} > \text{ and } \vec{P} = < \frac{\left|M_{i,f(i)}\right|}{Sum\{S\}} >$$

We perform this experiment with the LBJ and BART CRR systems on the ACE Phase 2 corpus. We illustrate the correlation results in Figure 1.

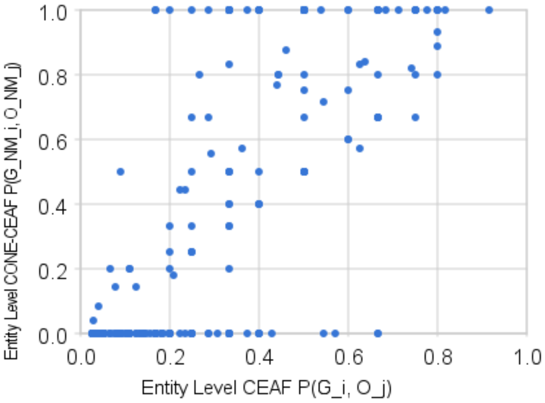

Figure 1. Correlation between $\vec{P}^{NM}$ and $\vec{P}$ - Entity Level CEAF Precision

From Figure 1, we can see that the two measures are highly correlated. In fact, we find that the Pearson's correlation coefficient (Soper et al., 1917; Cohen, 1988) is 0.73. The points lining up on the x-axis and y=1.0 represent very small equivalence classes and are a form of noise; their removal doesn't affect this coefficient. To show that this strong correlation is not a statistical anomaly, we also compute entity-level correlation using (G_i - $G^{NM}{}_i$, O_j - $O^{NM}{}_j$) and (G_i, O_j) instead of ($G^{NM}{}_i$, $O^{NM}{}_j$) and (G_i, O_j) and find that the coefficient drops to 0.03, which is obviously not correlated at all.

We now know $\vec{P}^{NM}$ and $\vec{P}$ are highly correlated. Assume the correlation is linear, with the following equation:

$$P_i = \alpha P^{NM}{}_i + \beta$$

where α and β are the linear regression parameters.

Thus

$$P = \sum_i P_i = \sum_i \left(\alpha P^{NM}{}_i + \beta\right) = n\alpha P^{NM} + n\beta$$

Here, *n* is the number of equivalence classes.

We conclude that the overall CEAF Precision and CONE CEAF Precision should be highly correlated too. We repeat this experiment with CEAF / CONE CEAF Recall, B^3 / CONE B^3 Precision and B^3 / CONE B^3 Recall and obtain similar results, allowing us to conclude that these sets of measures should also be highly correlated. We note here some generally accepted terminology regarding correlation: If two quantities have a Pearson's correlation coefficient greater than 0.7, they are considered "strongly correlated", if their correlation is between 0.5 and 0.7, they are considered "highly correlated", if it is between 0.3 and 0.5, they are considered "correlated", and otherwise they are considered "not correlated".

It is important to note that like all automatic evaluation metrics, CONE B^3 and CONE CEAF too can be easily 'cheated', e.g. a NE-CRR system that performs NER and named entity matching well but does not even detect and classify anaphora or nominal mentions would nonetheless score highly on these metrics. A possible solution to this problem would be to create gold standard annotations for a small subset of the data, call these annotations *G'*, and report two scores: B^3 / CEAF (*G'*), and CONE B^3 / CONE CEAF ($G^{NM\text{-}approx}$). Discrepancies between these two scores would enable the detection of such 'cheating'. A related point is that designers of NE-CRR systems should not optimize for CONE metrics alone, since by using $G^{NM\text{-}approx}$ (or G^{NM} where gold standard annotations are available), these metrics are obviously biased towards named mentions. This issue can also be addressed by having gold standard annotations *G'* for a small subset. One could then train a system by optimizing both B^3 / CEAF (*G'*) and CONE B^3 / CONE CEAF ($G^{NM\text{-}approx}$). This can be thought of as a form of semi-supervised learning, and may be useful in areas such as domain adaptation, where we could use some annotated test-set in a standard domain, e.g. newswire as the smaller set and an unlabeled large testset from some other domain, such as e-mail or biomedical documents. An interesting future direction is to monitor the effectiveness of our metrics over time. As co-reference resolution systems evolve in strength, our metrics might be less effective, however this could be a good indicator to discriminate on different subtasks the improvements gained by the co-reference resolution systems.

5 Experimental Results

We present experimental results in support of the validity and effectiveness of CONE metrics. As mentioned earlier, we used the following four publicly available CRR systems: UIUC's LBJ system (L), BART from JHU Summer Workshop (B), LingPipe from Alias-i (LP), and OpenNLP (OP) (Bengston and Roth, 2007; Versley et al., 2008; Baldridge and Torton, 2004; Baldwin and Carpenter, 2003). All these CRR systems perform Noun Phrase co-reference resolution (NP-CRR), not NE-CRR. So, we must first eliminate all equivalences classes that do not contain any named mentions. We do so using the SYNERGY NER system to separate named mentions from unnamed ones. Note that this must not be confused with the use of SYNERGY to produce G^{NM} and O^{NM} from *G* and *O* respectively. For that task, all equivalence classes in *G* and *O* already contain at least one named mention and we remove all unnamed mentions from each class. This process effectively converts the NP-CRR results of these systems into NE-CRR ones. We use the ACE Phase 2 NWIRE and ACE 2005 English datasets. We avoid using the ACE 2004 and MUC6 datasets because the UIUC LBJ system was trained on ACE 2004 (Bengston and Roth, 2008), while BART and LingPipe were trained on MUC6. There are 29 files in the test set of ACE Phrase 2 and 81 files in ACE 2005, summing up to 120 files with around 50,000 tokens with 5000 valid co-reference mentions. Tables 1 and 2 show the Pearson's correlation coefficients between CONE metric scores of the type Score(G^{NM}, O^{NM}) and standard metric scores of the type Score(*G*, *O*) for combinations of various CRR systems and datasets.

	B3/CONE B3			CEAF/CONE CEAF		
	P	**R**	**F1**	**P**	**R**	**F1**
L	0.82	0.71	0.7	0.81	0.71	0.77
B	0.85	0.5	0.66	0.71	0.61	0.68
LP	0.84	0.66	0.67	0.74	0.71	0.73
OP	0.31	0.57	0.61	0.79	0.72	0.79

Table 1. G^{NM}: Correlation on ACE Phase 2

	B3/CONE B3			CEAF/CONE CEAF		
	P	**R**	**F1**	**P**	**R**	**F1**
L	0.6	0.62	0.62	0.75	0.61	0.68
B	0.74	0.82	0.84	0.72	0.68	0.67
LP	0.91	0.65	0.73	0.44	0.57	0.53
OP	0.48	0.77	0.8	0.54	0.67	0.65

Table 2. G^{NM}: Correlation on ACE 2005

We observe from Tables 1 and 2 that CONE B^3 and CONE CEAF scores are highly correlated with B^3 and CEAF scores respectively, and this holds true for Precision, Recall and F1 scores, for all combinations of CRR systems and datasets. This justifies our assumption that a system's performance for the subtask of NMEG is a good predictor of its performance for the full task of NE-CRR. These correlation coefficients are graphically illustrated in Figures 2 and 3.

We now use our baseline named entity matching method to automatically generate estimated gold standard annotations $G^{NM\text{-}approx}$ and recalculate CONE CEAF and CONE B^3 scores using $G^{NM\text{-}approx}$ instead of G^{NM}. Tables 3 and 4 show the correlation coefficients between the new CONE scores and the standard metric scores.

	B3/CONE B3			CEAF/CONE CEAF		
	P	**R**	**F1**	**P**	**R**	**F1**
L	0.31	0.23	0.22	0.33	0.55	0.56
B	0.71	0.44	0.43	0.61	0.63	0.71
LP	0.57	0.43	0.49	0.36	0.25	0.31
OP	0.1	0.6	0.64	0.35	0.53	0.53

Table 3. $G^{NM\text{-}approx}$: Correlation on ACE Phase 2

	B3/CONE B3			CEAF/CONE CEAF		
	P	**R**	**F1**	**P**	**R**	**F1**
L	0.33	0.32	0.42	0.22	0.34	0.36
B	0.25	0.66	0.65	0.2	0.45	0.37
LP	0.19	0.33	0.34	0.77	0.68	0.72
OP	0.26	0.66	0.67	0.28	0.42	0.38

Table 4. $G^{NM\text{-}approx}$: Correlation on ACE Phase 2

We observe from Tables 3 and 4 that these correlation factors are encouraging, but not as good as those in Tables 1 and 2. All the corresponding CONE B^3 and CONE CEAF scores are correlated, but very few are highly correlated. We should note however that our baseline system to create $G^{NM\text{-}approx}$ uses relatively simple clustering methods and heuristics. It is easy to observe that a sophisticated named entity matching system would produce a $G^{NM\text{-}approx}$ that better approximates G^{NM} than our baseline method, and CONE B^3 and CONE CEAF scores calculated using this $G^{NM\text{-}approx}$ would be more correlated with standard B^3 and CEAF scores.

We note from the above results that correlations scores are very similar across different systems and datasets. In order to formalize this assertion, we calculate correlation scores in a system-independent and data-independent manner. We combine all the data points across all four different systems and plot them in Figure 2 and 3 for ACE Phase 2 NWIRE corpus and in Figure 4 and 5 for ACE 2005 corpus respectively. We illustrate only F1 scores; the results for precision and recall are similar.

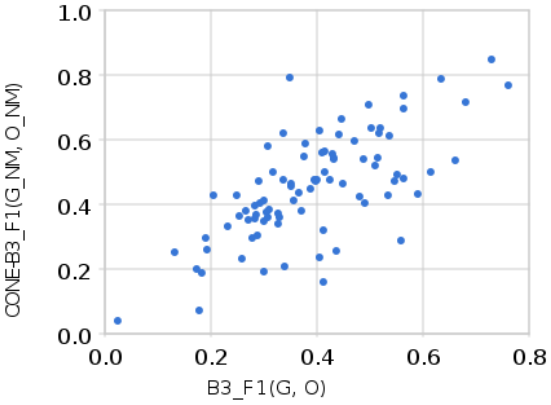

Figure 2. Correlation between B^3 F1 and CONE B^3 F1 for all systems on ACE 2

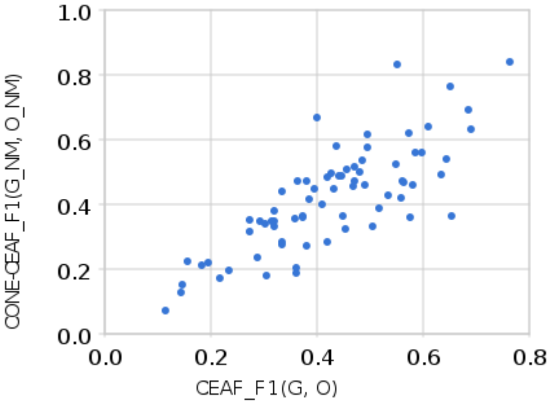

Figure 3. Correlation between CEAF F1 and CONE CEAF F1 for all systems on ACE 2

Figure 2 reflects a Pearson's correlation coefficient of 0.70, suggesting that all the B^3 F1 and CONE B^3 F1 scores for different systems are highly correlated and that CONE B^3 F1 does not bias towards any particular system. Figure 3 reflects a Pearson's correlation coefficient of 0.83, providing similar evidence for the system-independence of correlation between CEAF F1 and CONE CEAF F1 scores. Figures 4 and 5 corresponding to ACE 2005 reflect similar correlation coefficients of 0.89 and 0.82, and thus support the idea that the correlations between B^3 F1 and CONE B^3 F1, as well as between CEAF F1and CONE CEAF F1, are dataset-independent in addition to being system-independent.

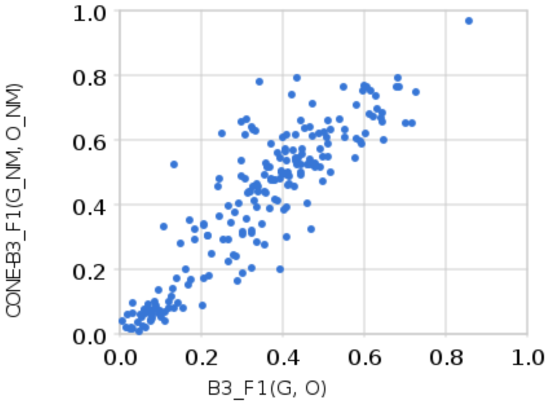

Figure 4. Correlation between B^3 F1 and CONE B^3 F1 for all systems on ACE 2005

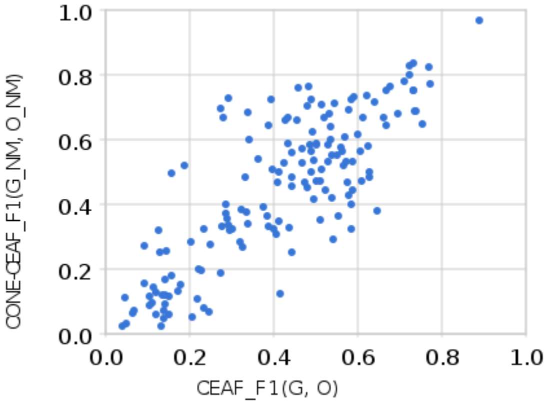

Figure 5. Correlation between CEAF F1 and CONE CEAF F1 for all systems on ACE 2005

6 Application and Discussion

To illustrate the applicability of CONE metrics, we consider the Enron e-mail corpus. It is of a different genre than the newswire corpora that CRR systems are usually trained on, and no CRR gold standard annotations exist for it. Consequently, no CRR systems have been evaluated on it so far. We used CONE B^3 and CONE CEAF to evaluate and compare the NE-CRR performance of various CRR systems on a subset of the Enron e-mail corpus (Klimt and Yang, 2004) that was cleaned and stripped of spam messages. We report the results in Table 5.

	CONE B^3			CONE CEAF		
	P	**R**	**F1**	**P**	**R**	**F1**
L	0.43	0.21	0.23	0.31	0.17	0.21
B	0.26	0.18	0.2	0.26	0.16	0.2
LP	0.61	0.51	0.53	0.58	0.53	0.54
OP	0.19	0.03	0.05	0.11	0.02	0.04

Table 5. $G^{NM\text{-}approx}$ Scores on Enron corpus

We find that LingPipe is the best of all the systems we considered, and LBJ is slightly ahead of BART in all measures. We suspect that since LingPipe is a commercial system, it may have extra training resources in the form of non-traditional corpora. Nevertheless, we believe our method is robust and scalable for large corpora without NE-CRR gold standard annotations.

7 Conclusion and Future Work

We propose the CONE B^3 and CONE CEAF metrics for automatic evaluation of Named Entity Co-reference Resolution (NE-CRR). These metrics measures a NE-CRR system's performance on the subtask of named mentions extraction and grouping (NMEG) and use it to estimate the system's performance on the full task of NE-CRR. We show that CONE B^3 and CONE CEAF scores of various systems across different datasets are strongly correlated with their standard B^3 and CEAF scores respectively. The advantage of CONE metrics compared to standard ones is that instead of the full gold standard data *G*, they only require a subset G^{NM} of named mentions which even if not available can be closely approximated by using a state-of-the-art NER system and clustering its results. Although we use a simple baseline algorithm for producing the approximate gold standard $G^{NM\text{-}approx}$, CONE B^3 and CONE CEAF scores of various systems obtained using this $G^{NM\text{-}approx}$ still prove to be correlated with their standard B^3 and CEAF scores obtained using the full gold standard *G*. CONE metrics thus reduce the need of expensive labeled corpora. We use CONE B^3 and CONE CEAF to evaluate the NE-CRR performance of various CRR systems on a subset of the Enron email corpus, for which no gold standard annotations exist and no such evaluations have been performed so far. In the future, we intend to use more sophisticated named entity matching schemes to produce better approximate gold standards $G^{NM\text{-}approx}$. We also intend to use the CONE metrics to evaluate NE-CRR systems on new datasets in domains such as chat, email, biomedical literature, etc. where very few corpora with gold standard annotations exist.

Acknowledgments
We would like to thank Prof. Ani Nenkova from the University of Pennsylvania for her talk about automatic evaluation for text summarization at the spring 2010 CMU LTI Colloquium and anonymous reviewers for insightful comments.

References

E. Agirre, L. Màrquez and R. Wicentowski, Eds. 2007. *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval).*

A. Bagga and B. Baldwin. 1998. Algorithms for Scoring Coreference Chains. *Proceedings of LREC Workshop on Linguistic Coreference*.

J. Baldridge and T. Morton. 2004. OpenNLP. http://opennlp.sourceforge.net/.

B. Baldwin and B. Carpenter. 2003. LingPipe. Alias-i.

E. Bengtson and D. Roth. 2008. Understanding the Value of Features for Coreference Resolution. *Proceedings of EMNLP*.

J. Cohen. 1988. *Statistical power analysis for the behavioral sciences*. (2nd ed.)

A.K. Elmagarmid, P.G. Ipeirotis and V.S. Verykios. 2007. Duplicate Record Detection: A Survey. *IEEE Transactions on Knowledge and Data Engineering*, v.19 n.1, 2007.

J.R. Finkel, T. Grenager, and C. Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. *Proceedings of ACL*.

B. Klimt and Y. Yang. 2004. The Enron corpus: A new dataset for email classification research. *Proceedings of ECML.*

H.W. Kuhn. 1955. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(83).

C. Lin and E. Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. *Proceedings of HLT-NAACL*.

C. Lin and F.J. Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. *Proceedings of ACL*.

A. Louis and A. Nenkova. 2009. Automatically Evaluating Content Selection in Summarization without Human Models. *Proceedings of EMNLP*, pages 306–314, Singapore, 6-7 August 2009.

X. Luo. 2005. On coreference resolution performance metrics. *Proceedings of EMNLP*.

MUC-6. 1995. *Proceedings of the Sixth Understanding Conference (MUC-6).*

J. Munkres. 1957. Algorithms for the assignment and transportation problems. *Journal of SIAM*, 5:32-38.

NIST. 2003. The ACE evaluation plan. www.nist.gov/speech/tests/ace/index.htm.

K Papineni, S Roukos, T Ward and W.J. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. *Proceedings of ACL*.

V. Pervouchine, H. Li and B. Lin. 2009. Transliteration alignment. *Proceedings of ACL*.

L. Ratinov and D. Roth. 2009. Design Challenges and Misconceptions in Named Entity Recognition. *Proceedings of CoNLL.*

R. Shah, B. Lin, A. Gershman and R. Frederking. 2010. SYNERGY: a named entity recognition system for resource-scarce languages such as Swahili using online machine translation. *Proceedings of LREC Workshop on African Language Technology*.

H.E. Soper, A.W. Young, B.M. Cave, A. Lee and K. Pearson. 1917. On the distribution of the correlation coefficient in small samples. Appendix II to the papers of "Student" and R. A. Fisher. A co-operative study. *Biometrika*, 11, 328-413.

V. Stoyanov, N. Gilbert, C. Cardie and E. Riloff. 2009. Conundrums in Noun Phrase Coreference Resolution: Making Sense of the State-of-the-Art. *Proceedings of ACL.*

Y. Versley, S.P. Ponzetto, M. Poesio, V. Eidelman, A. Jern, J. Smith, X. Yang and A. Moschitti. 2008. BART: A Modular Toolkit for Coreference Resolution. *Proceedings of EMNLP*.

M. Vilain, J. Burger, J. Aberdeen, D. Connolly and L. Hirschman. 1995. A model-theoretic coreference scoring scheme. *Proceedings of MUC 6.*