Which Coreference Evaluation Metric Do You Trust? A Proposal for a Link-based Entity Aware Metric

Nafise Sadat Moosavi and **Michael Strube**
Heidelberg Institute for Theoretical Studies gGmbH
Schloss-Wolfsbrunnenweg 35
69118 Heidelberg, Germany
{nafise.moosavi|michael.strube}@h-its.org

Abstract

Interpretability and discriminative power are the two most basic requirements for an evaluation metric. In this paper, we report the *mention identification effect* in the B^3, *CEAF*, and *BLANC* coreference evaluation metrics that makes it impossible to interpret their results properly. The only metric which is insensitive to this flaw is *MUC*, which, however, is known to be the least discriminative metric. It is a known fact that none of the current metrics are reliable. The common practice for ranking coreference resolvers is to use the average of three different metrics. However, one cannot expect to obtain a reliable score by averaging three unreliable metrics. We propose LEA, a Link-based Entity-Aware evaluation metric that is designed to overcome the shortcomings of the current evaluation metrics. *LEA* is available as branch `LEA-scorer` in the reference implementation of the official CoNLL scorer.

1 Introduction

There exists a variety of models (e.g. pairwise, entity-based, and ranking) and feature sets (e.g. string match, lexical, syntactic, and semantic) to be used in coreference resolution. There is no known formal way to prove which coreference model is superior to the others and which set of features is more beneficial/less useful in coreference resolution. The only way to compare different models, features or implementations of coreference resolvers is to compare the values of the existing coreference resolution evaluation metrics. By comparing the evaluation scores, we determine which system performs best, which model suits coreference resolution better, and which feature set is useful for improving the recall or precision of a coreference resolver. Therefore, evaluation metrics play an important role in the advancement of the underlying technology. It is imperative for the evaluation metrics to be reliable. However, it is not a trivial task to score output entities with various kinds of coreference errors.

Several evaluation metrics have been introduced for coreference resolution (Vilain et al., 1995; Bagga and Baldwin, 1998; Luo, 2005; Recasens and Hovy, 2011; Tuggener, 2014). Metrics that are being used widely are *MUC* (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), *CEAF* (Luo, 2005), and *BLANC* (Recasens and Hovy, 2011). There are known flaws for each of these metrics. Besides, the agreement between all these metrics is relatively low (Holen, 2013), and it is not clear which metric is the most reliable. The CoNLL-2011/2012 shared tasks (Pradhan et al., 2011; Pradhan et al., 2012) ranked participating systems using an *average* of three metrics, i.e. *MUC*, B^3, and *CEAF*, following a proposal by (Denis and Baldridge, 2009a). Averaging three unreliable scores does not result in a reliable one. Besides, when an average score is used for comparisons, it is not possible to analyse recall and precision to determine which output is more precise and which one covers more coreference information. This is indeed a requirement for coreference resolvers to be used in end-tasks. Therefore, averaging individual metrics is nothing but a compromise.

As mentioned by Luo (2005), interpretability and discriminative power are two basic requirements for a reasonable evaluation metric. In regard to the *interpretability* requirement a high score should indicate that the vast majority of coreference relations and entities are detected correctly. Similarly, a system that resolves none of the coreference relations or entities should get a zero score.

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 632–642,
Berlin, Germany, August 7-12, 2016. ©2016 Association for Computational Linguistics

	MUC			B^3			$CEAF_e$			BLANC		
	R	P	F_1	R	P	F_1	R	P	F_1	R	P	F_1
Base	69.31	76.23	72.60	55.83	66.07	60.52	54.88	**59.41**	57.05	57.46	65.77	61.31
More precise	69.31	**82.29**	**75.24**	53.94	**69.32**	60.67	50.92	55.85	53.27	53.60	66.45	59.25
Less precise^a	69.31	69.46	69.38	60.53	63.98	62.21	64.82	53.06	**58.35**	68.68	67.02	67.68
Less precise^b	69.31	74.70	71.90	**60.61**	69.14	**64.60**	**69.50**	47.61	56.51	**68.74**	**67.37**	**67.87**

Table 1: Counterintuitive values of B^3, CEAF and BLANC recall and precision.

An evaluation metric should also be *discriminative*. It should be able to discriminate between good and bad coreference decisions. In this paper, we report on a drawback for *B^3*, *CEAF*, and *BLANC* which violates the interpretability requirement. We also show that this flaw invalidates the recall/precision analysis of coreference outputs based on these three metrics. We then review the current evaluation metrics with their known flaws to explain why we cannot trust them and need a new reliable one. Finally, we propose *LEA*, a Link-based Entity Aware evaluation metric that is designed to overcome problems of the existing metrics. We have begun the process of integrating the *LEA* metric in the official CoNLL scorer[1] so as to continue the progress made in recent years to produce replicable evaluation metrics. In order to use the *LEA* metric, there is no additional requirement than that of the CoNLL scorer v8.01 [2].

2 The Mention Identification Effect

All the proposed evaluation metrics for coreference resolution use recall, precision and F_1 for reporting the performance of a coreference resolver. Recall is an indicator of the fraction of correct coreference information, i.e. coreference links or entities, that is resolved. Precision is an indicator of the fraction of resolved coreference information that is correct. F_1 is the weighted harmonic mean of recall and precision.

While we usually use F_1 for comparing coreference resolution systems, it is also important for the corresponding recall and precision values to be interpretable and discriminative. Coreference resolution is not an end-task itself but it is an important step toward text understanding. Depending on the task, recall or precision may be more important. For example, as Stuckhardt (2003) argues, a coreference resolver needs high precision to meet the specific requirements of text summarization and question answering.

In this section, we show that the recall and precision of the *B^3*, *CEAF* and *BLANC* metrics are neither interpretable nor reliable. We choose the output of the state-of-the-art coreference resolver of Wiseman et al. (2015) on the CoNLL 2012 English test set as the base output. The CoNLL 2012 English test set contains 222 documents (comprising 348 partially annotated sections). This test set contains 19,764 coreferring mentions that belong to 4,532 different entities.

In Table 1, *Base* represents the scores of (Wiseman et al., 2015) on the CoNLL 2012 test set. All reported scores in this paper are computed by the official CoNLL scorer v8.01 (Pradhan et al., 2014).

Assume $M_{k,r}$ is the set of mentions that exists in both key and response entities. Let $L_k(m)$ and $L_r(m)$ be the set of coreference links of mention m in the key and response entities, respectively. Mention m is an incorrectly resolved mention if $m \in M_{k,r}$ and $L_k(m) \cap L_r(m) = \emptyset$. Therefore, m is a coreferent mention that has at least one coreference link in the response entities. However, none of its detected coreference links in the response entities are correct.

By removing the incorrectly resolved mentions, the response entities will become more precise. The precision improves because the wrong links that are related to the incorrectly resolved mentions have been removed. Besides, the recall will not change because no correct coreference relations or entities have been added or removed.

We make the *Base* output more precise by removing all 1075 incorrectly resolved mentions from the response entities. The score for this more precise output is shown as *More precise* in Table 1. As can be seen, (1) recall changes for all the metrics except for *MUC*; (2) both *$CEAF_e$* recall and precision significantly decrease; and (3) *BLANC* recall notably decreases so that F_1 drops significantly in comparison to *Base*.

On the other hand, adding completely incorrect

[1] Currently available as branch `LEA-scorer` in `https://github.com/conll/reference-coreference-scorers`.

[2] *LEA* scores will be obtained by running the command `perl scorer.pl lea goldFile systemFile`.

entities to the response entities should not affect the recall and it should decrease the precision.

Assume $M_{d,k,\bar{r}}$ is the set of mentions of document d that exists in the key entities but is missing from the response entities. We can add completely incorrect entities to the *Base* output as follows: (1) By linking $m_1 \in M_{d,k,\bar{r}}$ to mention $m_2 \in M_{d,k,\bar{r}}$ that is non-coreferent with m_1. All the new wrong entities are of size two (*Less precise*^{a}). (2) By linking $m_1 \in M_{d,k,\bar{r}}$ to all mentions of $M_{d,k,\bar{r}}$ that are non-coreferent with m_1. In this case the new entities are larger but their number is smaller (*Less precise*^{b}). The number of new entities is 1350 and 283 for the first and second case, respectively. As can be seen from the results of Table 1, (1) recall changes for all metrics except for *MUC*; and (2) the *B*^{3}, *CEAF* and *BLANC* scores improve significantly over those of *Base* when the output is doubtlessly worse.

These experiments show that *B*^{3}, *CEAF* and *BLANC* are not reliable for recall-precision analysis. We refer to the problem that is causing these contradictory results as the **mention identification effect**.

3 Reasons for the Unreliable Results

In this section, we briefly give an overview of the common evaluation metrics for coreference resolution. We also discuss the shortcomings of each metric, including the **mention identification effect**, that may lead to counterintuitive and unreliable results. In all metrics, K is the key entity set and R is the response entity set.

3.1 MUC

MUC is the earliest systematic coreference evaluation metric and is introduced by Vilain et al. (1995). *MUC* is a link-based metric. It computes recall based on the minimum number of missing links in the response entities in comparison to the key entities. *MUC* recall is defined as:

$$\text{Recall} = \frac{\sum_{k_i \in K}(|k_i| - |p(k_i)|)}{\sum_{k_i \in K}(|k_i| - 1)}$$

where $p(k_i)$ is the set of partitions that is created by intersecting k_i with the corresponding response entities. *MUC* precision is computed by switching the role of the key and response entities.

It is not trivial to determine which evaluation metric discriminates coreference responses best. However, *MUC* is known to be the **least discriminative** coreference resolution metric (Bagga and Baldwin, 1998; Luo, 2005; Recasens and Hovy, 2011). The *MUC* evaluation is only based on the minimum number of missing/extra links in the response compared to the key entities. For instance, *MUC* does not differentiate whether an extra link merges two singletons or the two most prominent entities of the text. However, the latter error does more damage than the first one.

Another major problem with *MUC* is that it has an **incorrect preference in ranking** coreference outputs. *MUC* favors the outputs in which entities are over-merged (Luo, 2005). For instance, if we link all the key mentions of the CoNLL 2012 test set into a single response entity, the corresponding *MUC* scores, i.e. Recall=100, Precision=78.44 and F$_1$=87.91, will be all higher than those of the state-of-the-art system (*Base* in Table 1).

3.2 BCUBED

The *B*^{3} score is introduced by Bagga and Baldwin (1998). *B*^{3} is a mention-based metric, i.e., the overall recall/precision is computed based on the recall/precision of the individual mentions. For each mention m in the key entities, *B*^{3} recall considers the fraction of the correct mentions that are included in the response entity of m. *B*^{3} recall is computed as follows:

$$\text{Recall} = \frac{\sum_{k_i \in K} \sum_{r_j \in R} \frac{|k_i \cap r_j|^2}{|k_i|}}{\sum_{k_i \in K} |k_i|}$$

Similar to MUC, *B*^{3} precision is computed by switching the role of the key and response entities.

The **mention identification effect** arises in *B*^{3}, because *B*^{3} uses mentions instead of coreference relations to evaluate the response entities. Therefore, if a mention exists in a response entity, it is considered as a resolved mention regardless of whether it has a correct coreference relation in the response entity.

Luo (2005) argues that *B*^{3} leads to **counterintuitive results for boundary cases**: (1) consider a system that makes no decision and leaves every key mention as a singleton. *B*^{3} precision for this system is 100%. However, not all of the recognized system entities (i.e. singletons), or the detected coreference relations (i.e. every mention only coreferent with itself) are correct; (2) consider a system that merges all key mentions into a single entity. *B*^{3} recall for this system is 100%.

Luo (2005) interprets this recall as counterintuitive because the key entities have not been found in the response. The intuitiveness or counterintuitiveness of this recall value depends on the evaluator's point of view. From one point of view, all of the key mentions, that are supposed to be in the same entity, are indeed in the same entity.

Finally, as discussed by Luo and Pradhan (2016), *B*^{3} cannot properly handle **repeated mentions in the response entities**. If a gold mention is repeated in several response entities, *B*^{3} receives credit for all the repetitions. The repeated response mentions issue is not an imaginary problem (Luo and Pradhan, 2016). It can happen if system mentions are read from a parse tree where an NP node has a single child, a pronoun, and where both the nodes are considered as candidate mentions.

3.3 CEAF

The *CEAF* metric is introduced by Luo (2005). *CEAF*'s main assumption is that each key entity should only be mapped to one reference entity, and vice versa. *CEAF* uses a similarity measure (ϕ) to evaluate the similarity of two entities. It uses the Kuhn-Munkres algorithm to find the best one-to-one mapping of the key to the response entities (g^*) using the given similarity measure. Assuming K^* is the set of key entities that is included in the optimal mapping, recall is computed as:

$$\text{Recall} = \frac{\sum_{k_i \in K^*} \phi(k_i, g^*(k_i))}{\sum_{k_i \in K} \phi(k_i, k_i)} \qquad (1)$$

For computing *CEAF* precision, the denominator of Equation 1 is changed to $\sum_{R_i \in R} \phi(r_i, r_i)$. Based on ϕ, there are two variants of CEAF: (1) mention-based *CEAF* (*CEAF*$_m$), which computes the similarity as the number of common mentions between two entities, i.e. $\phi(k_i, r_j) = |k_i \cap r_j|$; and (2) entity-based *CEAF* (*CEAF*$_e$), in which $\phi(k_i, r_j) = \frac{2\times|k_i \cap r_j|}{|k_i|+|r_j|}$. The denominator of Equation 1 for *CEAF*$_e$ is the number of key entities.

Similar to *B*^{3}, the **mention identification effect** of *CEAF* is caused by both similarity measures of *CEAF* using the number of common mentions between two entities, i.e. $|k_i \cap r_j|$. In this way, even if the two mapped entities (k_i and r_j) have only one mention in common, *CEAF*$_m$ rewards recall and precision by $\frac{1}{\sum_{k_i}|k_i|}$ and $\frac{1}{\sum_{r_j}|r_j|}$, respectively. *CEAF*$_e$ rewards recall and precision by $\frac{2}{(|k_i|+|r_j|)\times|K|}$ and $\frac{2}{(|k_i|+|r_j|)\times|R|}$, respectively. If instead of the number of common mentions, we would use the number of common coreference links between two entities in both *CEAF*$_m$ and *CEAF*$_e$ similarity measures, this problem would be solved. However, even if we handle the mention identification effect by using coreference relations rather than mentions in the similarity measures, *CEAF* may still result in counterintuitive results. As mentioned by Denis and Baldridge (2009b), *CEAF* **ignores all correct decisions of unaligned response entities** that may lead to unreliable results. In order to illustrate this, we use a sample text from the CoNLL 2012 development set as an example (Figure 1). Gold mentions are enclosed in square brackets. Mentions with the same text are marked with different indices. The indices in parentheses denote to which key entity the mentions belong Consider cr_1 and cr_2 in Table 2, which are different responses for entity (1) of Figure 1. cr_1 resolves many coreference relations of entity (1). However, it misses that *they*$_1$ could refer to an entity which is already referred to by 'it'. Therefore cr_1 produces two entities instead of one because of this missing relation. On the other hand, cr_2 only recognizes half of the correct coreference relations of entity (1).

[The American administration]$_{(1)}$ committed a fatal mistake when [it$_1$]$_{(1)}$ [executed]$_{(2)}$ [this man]$_{(3)}$, in a way for which [it$_2$]$_{(1)}$ will pay a hefty price in the near future. [[His$_1$]$_{(3)}$ survival]$_{(4)}$ would have benefited [it$_3$]$_{(1)}$ much more than [[his$_2$]$_{(3)}$ execution]$_{(2)}$ if [they$_1$]$_{(1)}$ understood politics as [they$_2$]$_{(1)}$ should, because [[his$_3$]$_{(3)}$ survival]$_{(4)}$ could have been a card to threaten [the sectarians]$_{(5)}$ and keep [them$_1$]$_{(5)}$ as servants to [them$_1$]$_{(1)}$ and [their]$_{(1)}$ schemes.

Figure 1: Sample text from CoNLL 2012.

	Response entities
cr_1	r_1={the American administration, it$_1$, it$_2$, it$_3$} , r_2={they$_1$, they$_2$, them, their}
cr_2	r_1={the American administration, it$_1$, it$_2$, it$_3$}

Table 2: Different system outputs for Figure 1.

As can be seen from Table 3, CEAF prefers cr_2 over cr_1 even though cr_1 makes more correct decisions. CEAF only selects one of the output entities of cr_1 for giving credit to the correct decisions.

	MUC	B^3	CEAF$_m$	CEAF$_e$	BLANC
cr_1	92.30	66.66	50.00	44.44	60.00
cr_2	60.00	40.00	66.66	66.66	32.29

Table 3: F$_1$ scores for Table 2's response entities.

The other response entity is only used for penalizing the precision of cr_1. This counterintuitive result is only because of the stringent constraint of CEAF that the mapping of key to response entities should be one-to-one.

Another problem with $CEAF_e$, mentioned by Stoyanov et al. (2009), is that it **weights entities equally regardless of their sizes**. The system that does not detect entity (1), the most prominent entity of Figure 1, gets the same score as that of a system which does not detect entity (4) of size 2.

3.4 BLANC

BLANC (Recasens and Hovy, 2011; Luo et al., 2014) is a link-based metric that adapts the Rand index (Rand, 1971) to coreference resolution evaluation. Let C_k and C_r be the sets of coreference links in the key and response entities, respectively. Assume N_k and N_r are the sets of non-coreference links in the key and response entities, respectively. Recall and precision of coreference links are computed as:

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

Recall and precision of non-coreference links are computed as:

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

BLANC recall and precision are computed by averaging the recall and precision of coreference and non-coreference links, e.g. Recall= $\frac{R_c+R_n}{2}$.

The *BLANC* measure is the newest but the least popular metric for evaluating coreference resolvers. Because of considering non-coreferent relations, the **mention identification effect** affects *BLANC* most strongly. When the number of gold mentions that exist in the response entities is larger, the number of detected non-coreference links will also get larger. Therefore, it results in higher values for *BLANC* recall and precision ignoring whether those gold mentions are resolved.

4 LEA

In this section, we present our new evaluation metric, namely the Link-Based Entity-Aware metric (*LEA*). *LEA* is designed to overcome the shortcomings of the current evaluation metrics.

For each entity, *LEA* considers how important the entity is and how well it is resolved. Therefore, *LEA* evaluates a set of entities as follows:

$$\frac{\sum_{e_i \in E}(importance(e_i) \times resolution\text{-}score(e_i))}{\sum_{e_k \in E} importance(e_k)}$$

We consider the size of an entity as a measure of importance, i.e. $importance(e) = |e|$. Therefore, the more prominent entities of the text get higher importance values. However, according to the end-task or domain used, one can choose other importance measures based on factors besides e_i's size, e.g. e_i's entity type or e_i's mention types. For example, as suggested by Holen (2013), each mention carries different information values, and considering this information could benefit the quantitative evaluation of coreference resolution. The $importance$ measure of *LEA* is the appropriate place to incorporate this kind of information.

Entity e with n mentions has $link(e) = n \times (n-1)/2$ unique coreference links. The resolution score of key entity k_i is computed as the fraction of correctly resolved coreference links of k_i:

$$resolution\text{-}score(k_i) = \sum_{r_j \in R} \frac{link(k_i \cap r_j)}{link(k_i)}$$

For each k_i, *LEA* checks all the response entities to see whether they are partial matches for k_i. r_j is a partial match for k_i, if it contains at least one of the coreference links of k_i. Thus, if a response entity only contains one mention of k_i, it is not a partial mapping of k_i.

Having the definitions of $importance$ and $resolution\text{-}score$, *LEA* recall is computed as:

$$\text{Recall} = \frac{\sum_{k_i \in K}(|k_i| \times \sum_{r_j \in R} \frac{link(k_i \cap r_j)}{link(k_i)})}{\sum_{k_z \in K} |k_z|}$$

LEA precision is computed by switching the role of the key and response entities:

$$\text{Precision} = \frac{\sum_{r_i \in R}(|r_i| \times \sum_{k_j \in K} \frac{link(r_i \cap k_j)}{link(r_i)})}{\sum_{r_z \in R} |r_z|}$$

LEA handles singletons by self-links. A self-link is a link connecting a mention to itself. Self-links indicate that a mention is only coreferent with itself and not with other mentions. By considering self-links, the number of links in a singleton is one. If entity k_i is a singleton, $link(k_i \cap r_j)$ is one only if r_j is a singleton and contains the same mention as k_i.

In summary, *LEA* is a link-based metric with the following properties:

- *LEA* takes into account all coreference links instead of only extra/missing links. Therefore, it has more discriminative power than MUC.
- *LEA* evaluates resolved coreference relations instead of resolved mentions. *LEA* also does not rely on non-coreferent links in order to detect entity structures or singletons. Therefore, the **mention identification effect** does not apply to *LEA* recall and precision. As a result, one can trust *LEA* recall or precision.
- *LEA* allows one-to-many mappings of entities. Unlike *CEAF*, all correct coreference relations are rewarded by *LEA*. More splits (or similarly merges) in entity k_i result in a smaller $\sum_{r_j \in R} link(k_i \cap r_j)$. Therefore, splitting (merging) of an entity in several entities will be penalized implicitly in $resolution\text{-}score$.
- *LEA* takes the importance of missing/extra entities into account. Therefore, unlike *CEAF*$_e$, it differentiates between the outputs missing the most prominent and the smallest entities.
- *LEA* considers resolved coreference relations instead of resolved mentions. Therefore, the existence of repeated mentions in different response entities is not troublesome for *LEA*.

5 An Illustrative Example

In this section, we use the example from Pradhan et al. (2014) to show the process of computing the LEA scores. In this example, $K = \{k_1 = \{a, b, c\}, k_2 = \{d, e, f, g\}\}$ is the set of key entities and $R = \{r_1 = \{a, b\}, r_2 = \{c, d\}, r_3 = \{f, g, h, i\}\}$ is the set of response entities.

Here we assume that $importance$ corresponds to entity size. Hence, $importance(k_1) = 3$ and $importance(k_2) = 4$. The sets of coreference links in k_1 and k_2 are $\{ab, ac, bc\}$ and $\{de, df, dg, ef, eg, fg\}$, respectively. Therefore, $link(k_1) = 3$ and $link(k_2) = 6$. ab is the only common link between k_1 and r_1. There are no common links between k_1 and the two other response entities. Similarly, k_2 has one common link with r_3 and it has no common links with r_1 or r_2. Therefore, $resolution\text{-}score(k_1) = \frac{1+0+0}{3}$ and $resolution\text{-}score(k_2) = \frac{0+0+1}{6}$. As a result *LEA* recall is computed as:

$$\frac{\sum importance(k_i) \times resolution\text{-}score(k_i)}{\sum importance(k_j)}$$

$$= \frac{3 \times \frac{1}{3} + 4 \times \frac{1}{6}}{3 + 4} \approx 0.24$$

By changing the roles of key and response entities, *LEA* precision is computed as:

$$\frac{2 \times \frac{1+0}{1} + 2 \times \frac{0+0}{1} + 4 \times \frac{0+1}{6}}{2 + 2 + 4} \approx 0.33$$

6 Evaluation on Real Data

Table 4 shows the scores of the state-of-the-art coreference resolvers developed by Wiseman et al. (2015), Martschat and Strube (2015), and Peng et al. (2015). Clark and Manning (2015)'s resolver is also among the state-of-the-art systems but we did not have access to their output. Considering the average score of *MUC*, *B*^{3}, and *CEAF*$_e$, *Martschat*, and *Peng* perform equally. However, according to *LEA*, *Martschat* performs significantly better based on an approximate randomization test (Noreen, 1989). *CEAF*$_e$ also agrees with *LEA* for this ranking. However, *CEAF*$_e$ recall and precision are similar for *Peng* while based on *LEA*, *Peng*'s precision is marginally better than recall.

In addition to the state-of-the-art systems, we report the scores of boundary cases in the CoNLL 2012 test set in Table 4: (1) *sys-sing*: all system mentions as singletons; and (2) *sys-1ent*: all system mentions in a single entity.

Table 5 presents the evaluations of the participating systems in the CoNLL 2012 shared task (closed task with predicted mentions). The rankings are specified in parentheses. For the *LEA* rankings we also perform a significance test. The systems without significant differences have the same ranking. The main difference between the rankings of *avg.* and *LEA* is the rank of *xu*. Based on *LEA*, *xu* is significantly better than *chen* and *chunyuang*, while *avg.* ranks these two above *xu*. The recall values of *chen* and *chunyuang* for mention identification are 75.08 and 75.23, which are higher than those of the best performing systems, i.e 72.75 for *fernandes*, and 74.23 for *martschat*. *chen* and *chunyuang* include 1850 and 1735 gold mentions in their outputs that have not a single correct coreference link. On the other hand, the number of these gold mentions in *xu* is 757. Therefore, these different rankings could be a direct result of the **mention identification effect**.

Overall, using one reliable metric instead of an average score benefits us in two additional ways: (1) we can perform a significance test to check whether there is a meaningful difference, and (2) the recall and precision values are meaningful.

	MUC			B^3			CEAF$_e$			CoNLL	LEA		
	R	P	F$_1$	R	P	F$_1$	R	P	F$_1$	Avg. F$_1$	R	P	F$_1$
Wiseman	69.31	76.23	72.60	55.83	66.07	60.52	54.88	59.41	57.05	63.39	51.78	62.12	56.48
Martschat	68.55	77.22	72.63	54.64	66.78	60.11	52.85	60.30	56.33	63.02	50.64	62.87	56.10
Peng	69.54	75.80	72.53	56.91	65.40	60.86	55.49	55.98	55.73	63.04	51.91	58.97	55.21
sys-sing	0.00	0.00	0.00	19.72	39.05	26.20	50.32	4.99	9.08	11.76	0.00	0.00	0.00
sys-1ent	88.01	29.58	44.28	84.87	2.53	4.91	1.50	19.63	2.80	17.33	82.31	2.27	4.43

Table 4: Results on the CoNLL 2012 test set.

	MUC	B^3	$CEAF_m$	$CEAF_e$	BLANC	CoNLL avg.	LEA
fernandes	70.51 (1)	57.58 (1)	61.42	53.86 (1)	58.75	60.65 (1)	53.28 (1)
martschat	66.97 (3)	54.62 (2)	58.77	51.46 (2)	55.04	57.68 (2)	49.99 (2)
bjorkelund	67.58 (2)	54.47 (3)	58.19	50.21(3)	55.42	57.42 (3)	49.98 (2)
chang	66.38 (4)	52.99 (4)	57.10	48.94 (4)	53.86	56.10 (4)	48.50 (4)
chen	63.71 (7)	51.76 (5)	55.77	48.10 (5)	52.87	54.52 (5)	46.24 (6)
chunyuang	63.82 (6)	51.21 (6)	55.10	47.58 (6)	52.65	54.20 (6)	45.84 (6)
shou	62.91 (8)	49.44 (9)	53.16	46.66 (7)	50.44	53.00 (7)	43.97 (8)
yuan	62.55 (9)	50.11 (8)	54.53	45.99 (8)	52.10	52.88 (8)	44.76 (8)
xu	66.18 (5)	50.30 (7)	51.31	41.25 (11)	46.47	52.58 (9)	46.83 (5)
uryupina	60.89 (10)	46.24 (10)	49.31	42.93 (9)	46.04	50.02 (10)	41.15 (10)
songyang	59.83 (12)	45.90 (11)	49.58	42.36 (10)	45.10	49.36 (11)	41.25 (10)
zhekova	53.52 (13)	35.66 (13)	39.66	32.16 (12)	34.80	40.45 (12)	29.98 (12)
xinxin	48.27 (14)	35.73 (12)	37.99	31.90 (13)	36.54	38.63 (13)	29.22 (12)
li	50.84 (11)	32.29 (14)	36.28	25.21 (14)	31.85	36.11 (14)	27.32 (14)

Table 5: The results of the CoNLL 2012 shared task.

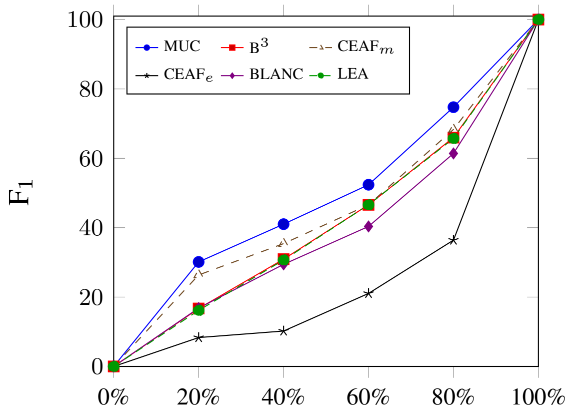

Figure 2: Resolved coreference links ratio without incorrect links.

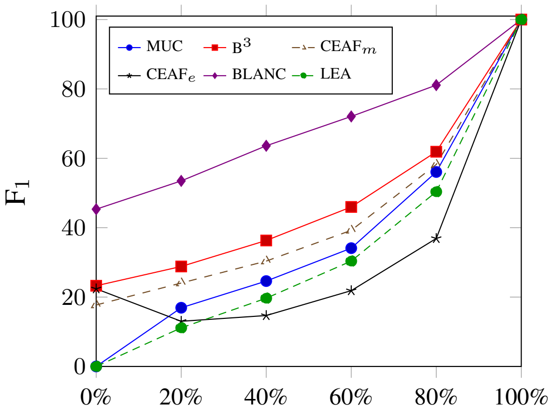

Figure 3: Resolved coreference links ratio in the presence of incorrect links.

7 Analysis

In this section we analyze the behavior of the evaluation metrics based on various coreference resolution errors. The set of key entities in all experiments contains: one entity of size 20, two entities of size 10, three entities of size 5, one entity of size 4, and ten entities of size 2.

7.1 Correct Links

We analyze different metrics based on the ratio of correctly resolved coreference links: (1) without wrong coreference links (Figure 2), and (2) with wrong coreference links (Figure 3). In the experiments of Figure 2, only mentions that are correctly resolved exist in the response. In Figure 3, apart from the mentions that are resolved correctly, other mentions are linked to at least one non-coreferent mention. Therefore, mention detection F$_1$ is always 100%.

The following observations can be drawn from these experiments: (1) *MUC* and *LEA* are the only measures which give a zero score to the response that contains no correct coreference relations; (2) in our experiments, *CEAF$_e$* shows an unreasonable drop when the correct link ratio changes from 0% to 20%; and (3), in Figure 2, the *BLANC*

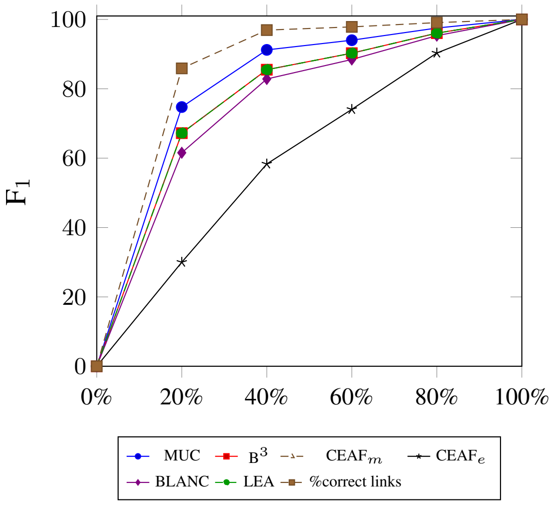

Figure 4: Resolving entities in decreasing order. F$_1$ of B^3, CEAF, and LEA are the same.

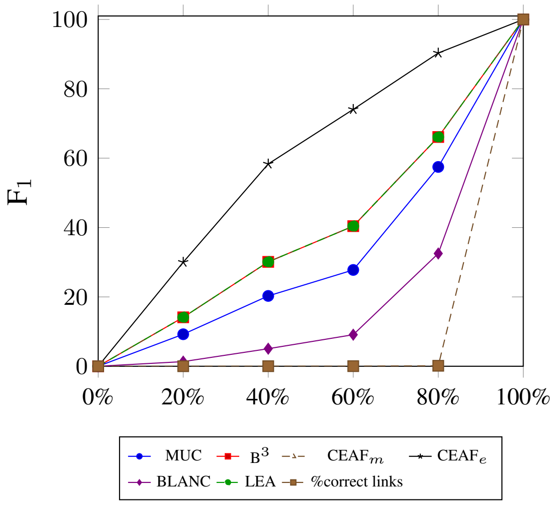

Figure 5: Resolving entities in increasing order. F$_1$ of B^3, CEAF, and LEA are the same.

F$_1$ values are less than or equal to those of B^3 and *LEA*. However, in Figure 3 that contains both coreferent and non-coreferent links, *BLANC* F$_1$ is at least 20% higher than that of other metrics.

7.2 Correct Entities

Apart from the correctly resolved links, a coreference metric should also take into account the resolved entities. In this section, we analyze the coreference resolution metrics based on the number and the size of the correctly resolved entities. In these experiments, each entity is either resolved completely, or all of its mentions are absent from the response. In Figure 4, the key entities are added to the response in decreasing order of their size. Figure 5 shows the experiments in which the entities are resolved in increasing order. The ratio of the correctly resolved coreference links is shown in both figures.

We can observe the following points from Figure 4 and Figure 5: (1) *CEAF$_e$* results in the same F$_1$ values regardless of the size of entities that are resolved or are missing; (2) B^3, *CEAF$_m$* and *LEA* result in the same F$_1$ values; and (3) *BLANC* is very sensitive to the total number of links.

7.3 Splitting/Merging Entities

The effect of splitting a single entity into two or more entities is studied in Figure 6. The overall effect of merging entities would be similar to that of splitting if the roles of the key and response entities change. In each experiment, only one key entity is split in a way that no singletons are created. For example, 18-2 in the horizontal axis indicates that an entity of size 20 is split into two entities of size 18 and 2.

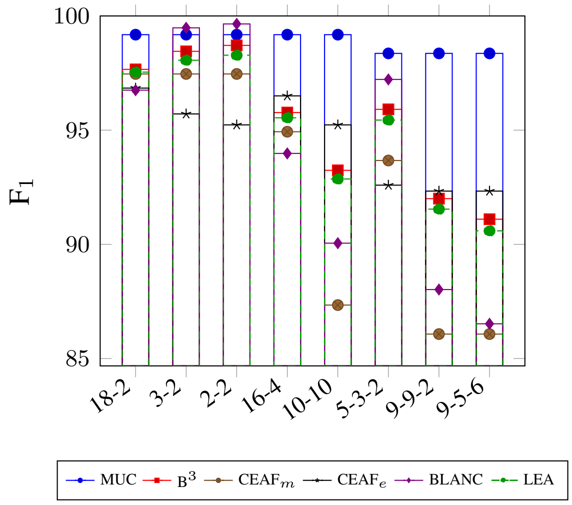

Figure 6: Effect of splitting entities.

The following observations can be drawn from Figure 6: (1) *MUC* only recognizes the number of splits regardless of the size of entities; (2) *CEAF$_e$* does not differentiate 2-2 from 10-10, and 9-9-2 from 9-5-6; and (3) the highest disagreement is for ranking different numbers of splits in entities with different sizes, i.e., B^3: 18-2>5-3-2>16-4, *BLANC*: 5-3-2>18-2>16-4, *CEAF*: 18-2>16-4>5-3-2, and *LEA*: 18-2>16-4>5-3-2. These are the cases that are even for humans hard to rank.

7.4 Extra/Missing Mentions

Figure 7 shows the effect of extra mentions, i.e. mentions that are not included in any key entity. If we change the roles of the key and response enti-

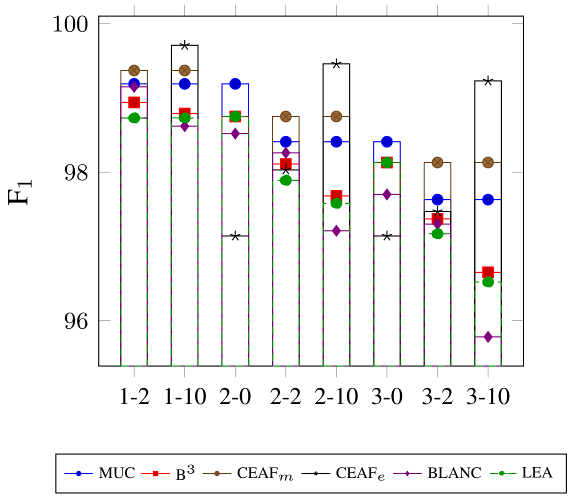

Figure 7: Effect of extra mentions.

ties, the overall effect of missing mentions would be similar. In the horizontal axis, the first number shows the number of extra mentions. The second number shows the size of the entity to which extra mentions are added. A zero entity size indicates the extra mentions are linked together.

The following points are worth noting from the results of Figure 7: (1) *MUC* and $CEAF_m$ are the least discriminative metrics when the system output includes extra mentions; (2) except for $CEAF_e$, other metrics rank 3-10 as the worst output;(3) $CEAF_e$ recognizes both 2-0 and 3-0 as the worst outputs. However, in these outputs the extra mentions are linked together and therefore no incorrect information is added to the correctly resolved entities; and (4) *LEA* is the only metric that recognizes error 2-0 is less harmful than 1-2 or 1-10. However, *LEA* does not discriminate the different outputs in which only one extra mention is added to an entity. If k extra mentions are added to an entity of size n, the corresponding resolution error multiplied by the importance of the response entity is $(n + k) \times (1 - \frac{n \times (n-1)}{(n+k) \times (n+k-1)})$. If $k = 1$, this equation is 2 regardless of n's value.

7.5 Mention Identification

The **mention identification effect** is shown in Figure 8. In all experiments, the number of correct coreference links is *zero*. The horizontal axis shows the mention identification accuracy in the system output. The F_1 of B^3, *CEAF* and *BLANC* in these experiments clearly contrast the interpretability requirement. A coreference resolver with a non-zero score should have resolved some of the coreference relations.

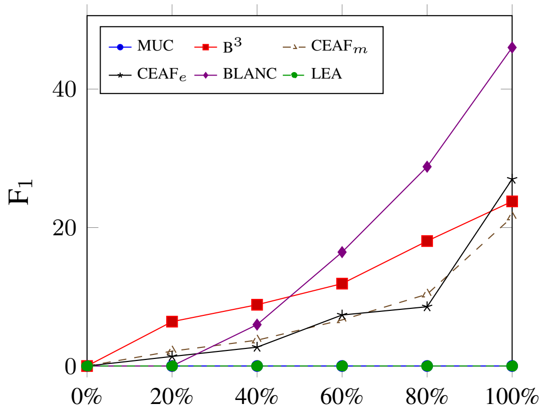

Figure 8: Effect of mention identification.

8 Conclusions

Current coreference resolution evaluation metrics have flaws which make them unreliable for comparing coreference resolvers. There is also a low agreement between the rankings of different metrics. The current solution is to use an average value of different metrics for comparisons. Averaging unreliable scores does not result in a reliable one. Indeed, recall and precision comparisons of coreference resolvers are not possible based on an average score. We first report the **mention identification effect** on B^3, *CEAF* and *BLANC* which causes these metrics to report misleading values. The only metric that is resistant to the mention identification effect is the least discriminative one, i.e. *MUC*. We introduce *LEA*, the Link-based Entity-Aware metric, as a new evaluation metric for coreference resolution. *LEA* is a simple intuitive metric that overcomes the drawbacks of the current metrics. It can be easily adapted for entity evaluation in different domains or applications in which entities with various attributes are of different importance.

Acknowledgments

The authors would like to thank Sameer Pradhan, Mark-Christoph Müller, Mohsen Mesgar and Sebastian Martschat for their helpful comments. We would also like to thank Sam Wiseman and Haoruo Peng for providing us with their coreference system outputs. This work has been funded by the Klaus Tschira Foundation, Heidelberg, Germany. The first author has been supported by a Heidelberg Institute for Theoretical Studies PhD. scholarship.

References

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *Proceedings of the 1st International Conference on Language Resources and Evaluation,* Granada, Spain, 28–30 May 1998, pages 563–566.

Kevin Clark and Christopher D. Manning. 2015. Entity-centric coreference resolution with model stacking. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* Beijing, China, 26–31 July 2015, pages 1405–1415.

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

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

Gordana Ilic Holen. 2013. Critical reflections on evaluation practices in coreference resolution. In *Proceedings of the 2013 NAACL HLT Student Research Workshop,* Atlanta, Georgia, 9-14 June 2013, pages 1–7.

Xiaoqiang Luo and Sameer Pradhan. 2016. Evaluation metrics. In M. Poesio, R. Stuckardt, and Y. Versley, editors, *Anaphora Resolution: Algorithms, Resources, and Applications*. Springer. To appear.

Xiaoqiang Luo, Sameer Pradhan, Marta Recasens, and Eduard Hovy. 2014. An extension of BLANC to system mentions. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 24–29, Baltimore, Maryland, June. Association for Computational Linguistics.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of the Human Language Technology Conference and the 2005 Conference on Empirical Methods in Natural Language Processing,* Vancouver, B.C., Canada, 6–8 October 2005, pages 25–32.

Sebastian Martschat and Michael Strube. 2015. Latent structures for coreference resolution. *Transactions of the Association for Computational Linguistics*, 3:405–418.

Eric W. Noreen. 1989. *Computer Intensive Methods for Hypothesis Testing: An Introduction.* Wiley, New York, N.Y.

Haoruo Peng, Kai-Wei Chang, and Dan Roth. 2015. A joint framework for coreference resolution and mention head detection. In *Proceedings of the 19th Conference on Computational Natural Language Learning,* Beijing, China, 30–31 July 2015, pages 12–21.

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 Shared Task of the 15th Conference on Computational Natural Language Learning,* Portland, Oreg., 23–24 June 2011, 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 the Shared Task of the 16th Conference on Computational Natural Language Learning,* Jeju Island, Korea, 12–14 July 2012, pages 1–40.

Sameer Pradhan, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng, and Michael Strube. 2014. Scoring coreference partitions of predicted mentions: A reference implementation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers),* Baltimore, Md., 22–27 June 2014, pages 30–35.

William R. Rand. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the 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(4):485–510.

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 Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing,* Singapore, 2–7 August 2009, pages 656–664.

Roland Stuckhardt. 2003. Coreference-based summarization and question answering: A case for high precision anaphor resolution. In *Proceedings of the 2003 International Symposium on Reference Resolution and Its Applications to Question Answering and Summarization,* Venice, Italy, 23–24 June 2003, pages 33–42.

Don Tuggener. 2014. Coreference resolution evaluation for higher level applications. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, Volume 2: Short Papers,* Gothenburg, Sweden, 26–30 April 2014, pages 231–235.

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 (MUC-6)*, pages 45–52, San Mateo, Cal. Morgan Kaufmann.

Sam Wiseman, Alexander M. Rush, Stuart Shieber, and Jason Weston. 2015. Learning anaphoricity and antecedent ranking features for coreference resolution. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* Beijing, China, 26–31 July 2015, pages 1416–1426.