Revisiting the Evaluation for Cross Document Event Coreference

Shyam Upadhyay Nitish Gupta Christos Christodoulopoulos Dan Roth
{upadhya3,ngupta19,christod,danr}@illinois.edu
University of Illinois at Urbana-Champaign, IL, USA

Abstract

Cross document event coreference (CDEC) is an important task that aims at aggregating event-related information across multiple documents. We revisit the evaluation for CDEC, and discover that past works have adopted different, often inconsistent, evaluation settings, which either overlook certain mistakes in coreference decisions, or make assumptions that simplify the coreference task considerably. We suggest a new evaluation methodology which overcomes these limitations, and allows for an accurate assessment of CDEC systems. Our new evaluation setting better reflects the corpus-wide information aggregation ability of CDEC systems by separating event-coreference decisions made across documents from those made within a document. In addition, we suggest a better baseline for the task and semi-automatically identify several inconsistent annotations in the evaluation dataset.

1 Introduction

Understanding events is crucial to natural language understanding and has applications ranging from question answering (Berant et al., 2014; Narayanan and Harabagiu, 2004), to causal reasoning (Do et al., 2011; Chambers and Jurafsky, 2008) to headline generation (Sun et al., 2015). The task of Cross Document Event Coreference (CDEC) determines if two event mentions (which belong to different documents) refer to the same event (Bagga and Baldwin, 1999; Bejan and Harabagiu, 2014). Figure 1 shows an example of two events whose mentions co-refer across 3 documents,

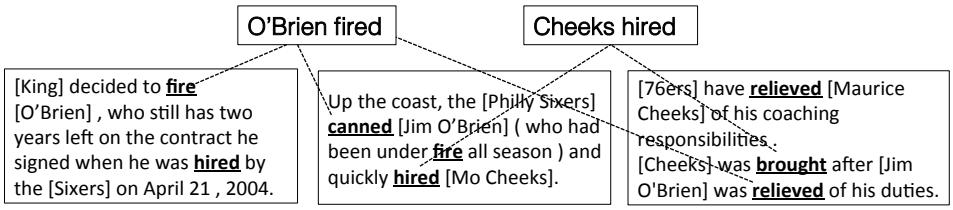


Figure 1: Event coreference across 3 documents. Event mentions are shown in **bold** and its participants are enclosed by brackets, []. Note that there are multiple **firing** events, but only a few co-refer.

An efficient CDEC system enables corpus-level aggregation of event attributes, which can prove valuable for tasks such as information extraction and aggregation (Humphreys et al., 1997; Zhang et al., 2015), topic detection and tracking (Allan et al., 1998), multi-document summarization (Daniel et al., 2003) and knowledge discovery (Mayfield et al., 2009).

In this work, we analyze whether existing CDEC evaluations reflect a system's ability to predict cross document links. Past works have adopted different evaluation methodologies which either make simplifying assumptions about the coreference task, such as ignoring singletons, or overlook certain coreference mistakes made by a CDEC system (as discussed in §4). Furthermore, under existing evaluations, a system which only predicts *within* document coreference links can score higher than a system which

predicts both within and cross document links. However, the latter system is clearly useful at aggregating information at corpus-level by finding across document links.

We address these issues by lifting the aforementioned assumptions, and proposing a new setting which enables accurate evaluation of the cross document coreference performance (§ 4.1). Lifting these assumptions also allowed us to semi-automatically identify several inconsistent annotations in the dataset. As the current evaluation dataset is the only available dataset with cross document coreference annotations, it is prudent to improve its annotation quality. We describe these annotation errors in § 7.3.

2 Related Work

Work on event coreference deals primarily with coreference within document, mostly building on insights gained from the entity coreference literature (Ng and Cardie, 2002; Bengtson and Roth, 2008; Lee et al., 2012; Peng et al., 2015). Recent approaches (Liu et al., 2014; Araki et al., 2014; Peng et al., 2016) have shown improvements in within document event coreference by exploiting event specific sub-structure (viz. sub-event or information propagation to arguments) and new event representations.

In comparison, cross document event coreference has been a less well-studied problem. Early work on cross document event coreference was done by Bagga and Baldwin (1999), who showed preliminary results on small exploratory datasets. To encourage research in this direction, Bejan and Harabagiu (2010) created the Event Coreference Bank (ECB), the first dataset with both within and across document event coreference annotations. ECB contained 482 documents, obtained from the Google News archive. Bejan and Harabagiu (2010) also showed encouraging results on ECB with several unsupervised Bayesian approaches. Later, ECB was augmented by Lee et al. (2012), to include entity level coreference annotations as well. On the new dataset, which they named EECB, they showed how entity and within document event coreference can benefit from making joint coreference decisions. Using EECB, Wolfe et al. (2015) formulated event coreference as an predicate alignment problem. Cybulska and Vossen (2014) noted that ECB and EECB did not have enough lexical diversity, thus oversimplifying the cross document coreference task. To get around this, they augmented the ECB corpus with 502 documents and released a larger corpus with event coreference annotations, named ECB+.

While there have been other works which use multi-modal supervision signals (Zhang et al., 2015) for CDEC, at present, ECB+ is the sole public dataset with cross document coreference annotations for events, leading to its popularity (Bejan and Harabagiu, 2014; Cybulska and Vossen, 2015; Yang et al., 2015). The most recent work using the ECB+ corpus is that of Yang et al. (2015), who developed a novel Bayesian clustering framework, which clusters event within and across documents, by modeling the clustering process as a Hierarchical Distance Dependent Chinese Restaurant Process (HDDCRP).

3 Cross Document Event Coreference

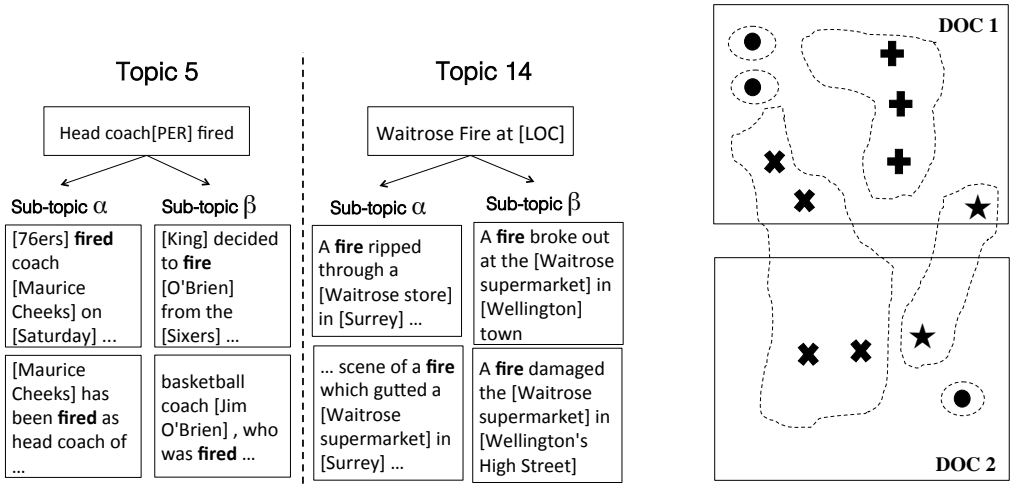
We view CDEC as a clustering task aimed at event-specific information aggregation. The clustering generated by a CDEC system allows one to examine all appearances of an event over a large corpus, shedding light on how the same event gets described in different documents.

We first describe our notation and the layout of the ECB+ dataset, and use Figures 2a and 2b to illustrate our definitions. Some statistics for the splits are shown in Table 1. We use the train, dev and test splits of Yang et al. (2015).

Event Mention A phrase which describes the action associated with an event. eg, in Figure 2a, *fired* is the event mention of the event "76ers fired coach Maurice Cheeks". We denote event mentions in **bold**.

Event Argument An event can have several arguments associated with it, describing its participants, location, or time of occurrence. eg., in Figure 2a, for the event "76ers fired coach Maurice Cheeks", "76ers" and "Maurice Cheek" are participant arguments and "Saturday" is the time argument. We collectively refer to these as the event arguments associated with the event mention (shown in [brackets]).

Event An event mention together with its arguments constitute an event, which describes an action and its arguments (location, time, participant etc.). An events in ECB+ can be in one of 3 categories:



(a) Layout of the ECB+ corpus.

(b) Event Types

Figure 2: (a) Documents in ECB+ are divided into topics and sub-topics, and coreference links can exist across documents in the same sub-topic. However, a coreference system is not aware of the topic (or sub-topic) partition at test time. (b) Coreference clusters are enclosed by a dotted line. Solid circles denote singleton mentions, solid crosses and stars denote cross document mentions, and solid plus denotes within document mentions.

(a) *Singleton Event* - an event with only one mention in the entire corpus. Shown as solid circles in Figure 2b, **(b)** *Within-Document (WD) Event* - an event with multiple mentions all of which appear in *one* document *only*. Shown as solid plus signs in Figure 2b, **(c)** *Cross-Document (CD) Event* - an event with multiple mentions spanning several documents. Shown as solid crosses and stars in Figure 2b.

Two events are *coreferent* if they represent the same situation. In particular, the text representation of the event should involve the same action and (some of the) arguments (Yang et al., 2015).

3.1 The ECB+ Corpus

In this section we discuss the specifics and layout of the ECB+ corpus.

Topics The documents in ECB+ are partitioned into several topics $\{T_1, T_2, ..., T_k\}$ each of which contains documents describing events of the same *event type*. For example, topic T_5 contains documents describing the firing of the head coach of a sports team. Figure 2a shows topics T_5 and T_{14} and the event types they described ("Head Coach [PER] fired" and "Waitrose Fire at [LOC]").

Sub-Topics A collection of documents that describe the same event within a topic, constitute a sub-topic. For example, the α sub-topic in topic 5 in Figure 2a, contains all documents describing the "firing of Maurice Cheeks", while those in the β sub-topic describe the "firing of O'Brien". Every topic T_i contains two sub-topics.[1] The documents in the β sub-topic were added by Cybulska and Vossen (2014) to increase the difficulty of the task, as a naive cross document event coreference system may incorrectly link a mention of the "O'Brien fire" event with a mention of the "Cheek fire" event.

3.2 Problem Formulation

Input: A collection of documents, each of which containing several event mentions.
Output: The system outputs an assignment of a cluster-id to each event mention, such that event mentions belonging to different documents but sharing the same cluster-id are coreferent.

Note that at test time, the CDEC system is not aware of the layout of the dataset so that it does not exploit the partitioning of documents into topics and sub-topics when making coreference decisions. As a result, a system can (incorrectly) link two event mentions across topics (or sub-topics), for which it should be appropriately penalized in evaluation. However, we will see in the next section that current evaluations do not meet this requirement.

[1] In principle, we can have more than 2 sub-topics per topic, but this does not occur in ECB+.

	Train	Dev	Test	Total
docs	462	73	447	982
topics	20	3	20	43
sub-topics	40	6	40	86
ev. mentions	5443	608	8951	14874
singleton	1866	167	5572	7605
WD	23	0	89	112
CD	3554	441	3290	7285
WD chains	2502	316	2138	4956
CD chains	691	47	479	1217

Table 1: Statistics of the ECB+ dataset. WD, CD refer to within document and cross document mentions respectively. WD (CD) Chains count the number of mention clusters that refer to the same event within (across) document(s). See text for details.

4 Evaluation Settings

All evaluation settings transform the cross document coreference problem to a within document coreference problem, by merging documents to create meta-document(s), such that cross document event coreference chains correspond to within document chains in the meta-document(s). We first revisit the evaluation settings that have been used in previous work. All these rely on the layout of the corpus, which affects their strictness.

Bejan and Harabagiu (2014) (B&H) In this case, the documents belonging to each topic are merged together to form a single meta-document $M(T_i)$. In this way, we have k meta-documents, one for each topic. Each meta-document is then evaluated separately for within document coreference, and the scores are aggregated by taking the micro-average. This evaluation is also followed by other works (Cybulska and Vossen, 2014; Cybulska and Vossen, 2015).

In this setting each topic's meta-document is evaluated in isolation, assuming that there were no coreference links made by a system across topics. However, a system can potentially link mentions across topics, as the input to the system does not describe the topic (or sub-topic) layout. As a result, this evaluation is oblivious to incorrect coreference made across topics. For example, it will ignore a (incorrect) link between the fire event mentions in topic T_5 and topic T_{14} in Figure 2a.

Yang et al. (2015) (YCF) In their case, all documents which belong to the same sub-topic (say α) of topic T_i are merged to create a meta-document $M(T_i, \alpha)$ for each sub-topic. Each such meta-document is then evaluated separately for within document coreference. In addition, all singleton event mentions are removed from the keys of all documents. [2]

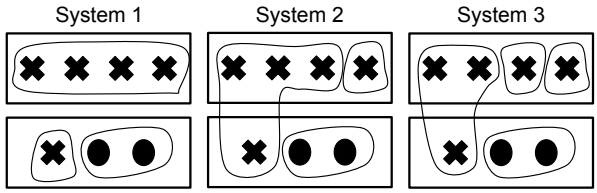
This setting implicitly makes two simplifying assumptions. First, it ignores singletons (referred to as I.S.) during evaluation, this making the coreference task considerably simpler (singletons constitute over 60% of the test mentions) as the system does not get penalized for making incorrect links to mentions of singleton events. Furthermore, by creating separate meta-document per sub-topic, this setting disregards the inclusion of documents from the β sub-topic into the corpus (see § 3), since incorrect coreference links across the sub-topics will not be penalized. This limitation (referred in Table 2 as S.T.) is similar, but more lenient, to the one in the B&H setting. (see § 7).

4.1 Our Proposals

We now discuss our proposed evaluation methodologies which address the limitations discussed above.

Proposal 1 (SIMPLE-CDEC) In this case, we merge all documents, across topics, into a single meta-document M. Unlike (Yang et al., 2015), we consider all event mentions (including singletons) in the corpus, and do not create a separate meta-document for each sub-topic, making our evaluation closer to what Cybulska and Vossen (2014) envisioned.

[2] We confirmed the YCF setting through personal communication with the authors and also examining the key and response files they provided.



(a) Enclosing rectangles denote documents, solid crosses (and circles) denote coreferent mentions of the same event (gold clustering). Polygon regions denote system response.

	MUC	B^3	$CEAF_e$	BLANC
System 1	88.8	87.0	75.5	80.5
System 2	88.8	87.0	75.5	80.5
System 3	74.9	74.9	58.3	63.7

(b) Performance of different systems as evaluated under SIMPLE-CDEC. Systems performing WD coreference are awarded equally or more than system performing CD coreference.

Figure 3: (a) Outputs of three different systems. System 1 only performs within-document (WD) coreference. System 2 and 3 performs cross-document(CD) coreference and correctly identify the cross document link of the solid cross event, but make mistake in the WD coreference. (b) SIMPLE-CDEC evaluation awards System 1 and System 2 same scores, and System 3 lower scores. Note that this issue is not limited to SIMPLE-CDEC, but also B&H and YCF.

SIMPLE-CDEC fulfills the requirement that the evaluation of a CDEC system should not be aware of the corpus layout into topics (as in ECB+) and all output links should be appropriately evaluated, unlike what is currently done in B&H and YCF. [3]

However, the SIMPLE-CDEC setting (just like B&H and YCF) does not distinguish the cross document performance from the within document performance. For example, in Figure 3a, a system which performs only within document coreference can get the same score as a system that performs both within and across document coreference (see Table 3b). However, it is desirable to have an evaluation setting that facilitates this distinction, and focuses on the ability of a CDEC system to aggregate information *across* documents in a large corpus, that is, evaluate on purely cross document coreference links. A CDEC system which performs within document event coreference only is not fully exploiting the large corpus and does not discover links between documents.

Proposal 2 (PURE-CDEC) In this setting, we first preprocess each document so that all within document mentions of the same event are reduced to a single meta-mention in that document. In this way, we discount the within document coreference links and focus on the cross document coreference links. Then, we follow the SIMPLE-CDEC setting to evaluate a single meta-document M generated from the preprocessed system response and the gold key documents.

5 Evaluation Metrics

We briefly describe the coreference evaluation metrics that are used in our experiments.

MUC (Vilain et al., 1995) A link-level metric, MUC counts the minimum number of edge-insertions or edge-deletions required to obtain the gold clustering (key) from the predicted clustering (response). A known limitation of MUC is that it does not reward a system for correctly identifying singletons.

B^3 (Bagga and Baldwin, 1998) A mention-level metric, B^3 calculates precision and recall for each mention by measuring the proportion of overlap between the predicted and gold coreference chains. The final score is an average over the scores for all mentions. Although it overcomes the limitations of MUC, it uses mentions of the same entity (coreference chain) more than once.

$CEAF_e$ (Luo, 2005) An entity-level metric which first finds an optimal alignment between entities in the key to the entities in the response by maximizing an entity similarity objective. This alignment is then used to calculate the CEAF precision and recall.

Blanc (Luo et al., 2014) First described in (Recasens and Hovy, 2011) and later extended in (Luo et al., 2014). Blanc is based on the Rand Index (Rand, 1971), and computes two F-scores, one evaluating the quality of coreference decisions (Pairwise) and another evaluating the quality of the non-coreference

[3] While evaluation should not rely on the corpus layout, we do not in any way suggest that this is the ideal approach for performing coreference. Indeed, considering coreference decisions over the entire corpus will be prohibitively expensive.

decisions (Pairwise-Negative). The Pairwise (PW) and the Pairwise-Negative (PWN) F-scores are then averaged to compute the final Blanc F-score.

Each of the above metrics have some drawbacks. B^3 and CEAF scores rapidly approach 100 if many singletons are present, and the same can drive the Blanc-PWN score to dominate the Blanc-PW score (Recasens and Hovy, 2011). This is why the entity coreference literature reports the CoNLL average, which is average of MUC, B^3 and $CEAF_e$ F-scores (Pradhan et al., 2011). Following previous work, we report CoNLL F1 and also the F-score averaged across all metrics.

Coreference Scorers For running evaluations under the YCF and B&H setting we use the standard within-document coreference scorer of Pradhan et al. (2014). However, in the SIMPLE-CDEC and PURE-CDEC setting, creating a single meta-document for the entire test split leads to a document with over 8000 mentions, which causes runtime issues for metrics like $CEAF_e$ and Blanc with Pradhan et al. (2014)'s scorer.[4] We use Hachey et al. (2014)'s scorer instead, which provides more efficient implementations.[5]

6 Baselines

We evaluate the following event coreference baseline models under the different evaluation settings,

Lemma In this model two events are coreferent if the head lemmas of their event mentions match.

Lemma-WD In this model we consider two event mentions coreferent if their head lemmas match and they belong to the same document. This is the within document variant of the Lemma model.

Lemma-δ In this model two event mentions are considered coreferent only if their head lemmas match and the *tf-idf* document similarity of the documents containing them exceeds a threshold (we use $\delta = 0.3$ after tuning on dev). For $\delta = 0$ this reduces to the Lemma baseline.

Supervised Agglomerative Clustering (SAC) Performs greedy agglomerative clustering, using a learned pairwise classifier to score if two cluster of mentions are coreferent. Inter-cluster similarity score is computed as the average of the similarity score of their mentions. Like Lemma-δ, two clusters are considered for linking if the document similarity exceeds a threshold (we use 0.3 as above).

We re-implemented the pairwise classifier of Yang et al. (2015) using the same feature set. Mention heads were found using dependency parses obtained by Stanford CoreNLP (Manning et al., 2014). We identify the Framenet (Baker et al., 1998) frame evoked by an event mention, we use the SEMAFOR (Das and Smith, 2011). For identifying the arguments for an event mention we use Illinois-SRL (Punyakanok et al., 2008). The pairwise classifier was trained using the Illinois-SL package (Chang et al., 2015).

7 Experiments

We first show the limitations of using current evaluation settings by showing how they can produce a wide range of results for the same baseline system. Next, we show the value of isolating the performance on predicting cross document coreference links. Finally, we assess the quality of the dataset and describe how we semi-automatically detected different types of annotation errors. Since we focus here on evaluation and not on developing an end-to-end CDEC system, we use the gold event mentions provided.

7.1 Comparing Evaluation Settings

We evaluate the lemma baseline under B&H, YCF, and SIMPLE-CDEC as described in §4. We aim to show that the same baseline approach can achieve highly variable results depending on the evaluation setting employed. We also isolate the effect of the two assumptions made by the YCF setting – first by creating a meta-document per sub-topic (S.T. in Table 2), and then by ignoring singletons (I.S. in Table 2) in SIMPLE-CDEC setting.

The results are shown in Table 2. Compared to SIMPLE-CDEC, all other settings assign the lemma baseline unrealistically high scores. B&H does not penalize links predicted across topics and therefore the F-score of the lemma baseline across all metrics increases. In addition to above, S.T. does not

[4] The $CEAF_e$ and Blanc evaluation did not finish after > 30 hours.

[5] Available at `github.com/wikilinks/neleval`.

Eval. Setting	MUC			B^3			$CEAF_e$			Blanc			CoNLL	Avg of 4
	P	R	F	P	R	F	P	R	F	PW	PWN	F	avg.	avg.
Settings that include singletons.														
SIMPLE-CDEC	30.5	75.7	43.4	28.4	81.9	42.2	65.6	18.6	29.0	11.1	**98.4**	54.8	38.2	42.3
B&H	37.4	75.3	50.0	49.4	81.6	61.5	39.9	**70.8**	51.0	30.5	93.4	62.0	54.2	56.1
S.T.	**40.9**	**75.9**	**53.2**	**58.8**	**82.7**	**68.7**	**71.3**	45.7	**55.7**	**37.0**	95.4	**66.2**	**59.2**	**61.0**
Settings that ignore singletons.														
I.S.	79.5	**76.1**	77.8	50.7	**54.0**	52.3	**39.9**	46.7	43.1	31.0	98.4	64.7	57.7	59.5
YCF (=IS+ST)	**94.5**	75.8	**84.2**	**92.0**	53.6	**67.8**	36.2	**75.2**	**48.9**	**53.3**	**98.7**	**76.0**	**67.0**	**69.2**

Table 2: **Evaluating the lemma baseline in different settings**. S.T. creates a separate meta-document for each sub-topic, while I.S. ignores singleton event mentions during evaluation. Note that the evaluation becomes more lenient from top to bottom, with SIMPLE-CDEC being the most strict. Best values in each column under different settings are shown in **bold**.

Baseline	MUC			B^3			$CEAF_e$			Blanc			CoNLL	Avg of 4
	P	R	F	P	R	F	P	R	F	PW	PWN	F	avg.	avg.
SIMPLE-CDEC														
Lemma-WD	38.0	20.4	26.5	**88.7**	68.4	**77.2**	67.5	**80.8**	**73.6**	5.3	**98.5**	51.9	59.1	57.3
Lemma	30.5	**75.7**	43.4	28.4	**81.9**	42.2	65.6	18.6	29.0	11.1	98.4	54.8	38.2	42.3
Lemma-δ	40.9	72.5	**52.3**	59.0	81.1	68.3	**73.6**	45.5	56.2	**32.8**	**98.5**	**65.6**	58.9	60.6
SAC	**44.2**	52.9	48.2	75.2	76.0	75.6	70.5	62.6	66.3	28.6	**98.5**	63.5	**63.4**	**63.4**
PURE-CDEC														
Lemma-WD	0.0	0.0	0.0	**90.1**	65.6	**75.9**	67.0	**80.2**	**73.0**	0.0	76.2	38.1	49.6	46.8
Lemma	18.7	**62.7**	28.8	27.2	**74.8**	39.9	65.7	18.6	29.0	7.2	76.2	41.7	32.6	34.9
Lemma-δ	29.4	42.4	34.7	68.6	71.6	70.1	70.6	56.8	62.9	**26.5**	76.2	51.3	53.0	52.6
SAC	**30.5**	42.0	**35.3**	70.6	74.5	72.5	**71.2**	63.2	67.0	25.3	**79.0**	**52.2**	**58.3**	**56.7**

Table 3: **Comparing the baseline approaches under** SIMPLE-CDEC **and** PURE-CDEC. Best values in each column under different settings are shown in **bold**.

penalize cross sub-topic predictions as well, which boosts the performance further. This confirms that settings like B&H do not penalize certain incorrect coreference links by exploiting the corpus layout.

Next, we compare the settings that ignore singleton event mentions (which constitute over 60% of the test data) during evaluation. When ignoring singletons (I.S.), the F-score for all metrics improves compared to SIMPLE-CDEC. Under I.S., incorrect coreference links to singleton events are discounted, resulting in high scores. Finally, with YCF, which combines the assumptions of I.S. and S.T., the scores again improve dramatically.

This experiment clearly shows that the assumptions made by B&H and YCF lead to lenient evaluations. SIMPLE-CDEC is a more appropriate evaluation setting since it remains unaware of the corpus layout and correctly penalizes all incorrect coreference link predictions when evaluating a CDEC system.

7.2 Understanding Cross Document Coreference Performance

In this experiment we evaluate all the baselines using SIMPLE-CDEC and PURE-CDEC. The aim is to show that if we do not evaluate the cross document coreference separately, a system can exploit the evaluation by only focusing on correctly predicting within document coreference links.

The results are shown in Table 3. Lemma-WD predicts lots of singletons as it only links mention in the same document. This results in high B^3 and CEAF scores in both SIMPLE-CDEC and PURE-CDEC since over 60% event mentions in the test data are singletons.

For SIMPLE-CDEC, the lemma-WD baseline performs surprisingly well, second to the SAC model. All metrics reward Lemma-WD for identifying the within document links, which leads to the high CoNLL score. Surprisingly, its CoNLL score is slightly higher than Lemma-δ, which is a model which does both within and across linking. This is because the latter may incorrectly link some singletons across documents. Note that this behavior is not exclusive to the SIMPLE-CDEC – a system which exclusively

predicts within document links can achieve such high scores under the B&H and YCF settings.

However, we can get a clearer assessment of cross document coreference performance by evaluating the models under the PURE-CDEC setting. As this evaluation only rewards finding correct cross document links, Lemma, Lemma-δ and SAC model are the only models which get non-zero MUC and Blanc-PW scores. As a result, the Lemma-WD average F-score drops considerably (by almost 10%) in comparison to the Lemma-δ and SAC model (around 6%), revealing that these models indeed do better cross document coreference. It is evident that to accurately assess the cross document coreference performance, we should report *both* PURE-CDEC and SIMPLE-CDEC results.

It should be noted that Lemma performs worse than Lemma-WD because it makes incorrect cross document links, due to the naive nature of the lemma match. On the other hand, Lemma-WD does not make *any* across document links, avoiding incurring these penalties. It gets rewarded for "identifying" singletons as described earlier in SIMPLE-CDEC. This conservative nature of Lemma-WD gets better average scores than Lemma, even in the PURE-CDEC setting. Overall, SAC and Lemma-δ are the two best models in Table 3.

Choice of Baseline Besides the aforementioned insights, Table 3 offers a better choice of baselines. Bejan and Harabagiu (2014) and Yang et al. (2015) claimed that Lemma is a strong baseline for CDEC. We believe that this claim held in these works only due to the lenient evaluation settings of B&H and YCF, which did not appropriately penalize the incorrect across topic (and sub-topic) links made by the Lemma baseline. However, the SIMPLE-CDEC and PURE-CDEC evaluations show that Lemma-δ is a stronger baseline. For future comparisons, using Lemma-δ as a baseline is more appropriate.

7.3 The Annotation Quality of ECB+

Evaluating with singletons also helped in discovering annotation errors in the dataset. In addition to identifying annotation errors, as described below, we found that several documents were partially annotated,[6] which is consistent with a similar observation made in (Liu et al., 2014).

We used the approach of Goldberg and Elhadad (2007) to semi-automatically detect annotation errors, by training an anchored SVM. First, for each pair of mention (m_i, m_j) in the training data, we added a unique anchor feature a_{ij}, thus making the data linearly separable. Next, we trained a SVM classifier on all of the data with a high penalty parameter C. The classifier uses the anchor features to memorize the hard to classify examples, which are either genuine hard coreference pairs, or incorrect annotations. By thresholding the features weights for the anchor features $|a_{ij}| > \delta$ (we use $\delta = 0.95$), we generated a short-list of these hard cases, which we then examined by a annotator for mistakes. The errors we found can be categorized as one of:

Missing Singleton-to-Singleton Link Two gold event mentions which should have been marked as coreferent, but were marked as singletons. For example:

Aceh was hit extremely hard by the massive Boxing Day **earthquake** and **tsunami** in 2004, killing 170,000 people
A massive **quake** struck off Aceh in 2004 , sparking a **tsunami** that killed 170,000 people ...

Both mention pairs (*earthquake*,*quake*) and (*tsunami*,*tsunami*) refer to the same event, but their coreference links were missing in the annotation. Both the enclosing documents belong to the same topic.

Missing Singleton-to-Cluster Link A singleton event mention which should have been linked to an existing cluster of mentions describing the same event.

LaRue , a Mississippi oil heir who became the first person found guilty of participating in the **Watergate** coverup
... strategy for capturing Southern votes and then a significant participant in the **Watergate scandal** .

The *Watergate scandal* mention is marked as singleton. However, the watergate scandal event appears in several other documents to which the *Watergate* mention is marked as being coreferent. Again, both the mentions belong to documents in the same topic.

We found over 300 such annotation errors which were incorrectly not linking singleton mentions. The list of errors detected is available at `http://cogcomp.cs.illinois.edu/page/publication_view/801`.

[6] For example, Only 5 sentences out of 40 in a document were marked with events.

8 Conclusion

Accurate evaluation and high-quality annotations are crucial to our ability to measure progress in any task. We showed that past work for CDEC have resorted to widely different evaluation approaches, making several implicit assumptions, which simplify the coreference task and lead to overlooking coreference mistakes. In particular, as we showed, excluding singleton mentions from the evaluation does not seem justified. Furthermore, current evaluation methods heavily rely on the corpus being organized into topics and sub-topics, but this may not always be available, especially for evaluation corpora. To accurately measure CDEC performance, it is necessary to drop these assumptions. We recommend that future evaluations report results using both SIMPLE-CDEC and PURE-CDEC settings.

Beyond these assumptions, the annotations in the current dataset are incomplete in several respects. Indeed, we showed that over 300 annotation errors in the dataset can be detected semi-automatically, and also noted that many documents were partially annotated. As this dataset is presently the only dataset with cross document coreference annotations for events, such annotation errors make evaluation difficult. We believe that to correctly evaluate this task and make progress, efforts must be made to create thoroughly annotated datasets with high quality annotations.

Acknowledgements

Thanks to Bishan Yang, Piek Vossen and Joel Nothman for answering questions and sharing system outputs. This work was supported by Contract HR0011-15-2-0025 with the US Defense Advanced Research Projects Agency (DARPA). Approved for Public Release, Distribution Unlimited. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

References

James Allan, Jaime G Carbonell, George Doddington, Jonathan Yamron, and Yiming Yang. 1998. Topic detection and tracking pilot study final report.

Jun Araki, Zhengzhong Liu, Eduard H Hovy, and Teruko Mitamura. 2014. Detecting subevent structure for event coreference resolution. In *LREC*.

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*.

Amit Bagga and Breck Baldwin. 1999. Cross-document event coreference: Annotations, experiments, and observations. In *the Workshop on Coreference and its Applications*.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *ACL-COLING*.

Cosmin Bejan and Sanda Harabagiu. 2010. Unsupervised event coreference resolution with rich linguistic features. In *ACL*.

Cosmin Adrian Bejan and Sanda Harabagiu. 2014. Unsupervised event coreference resolution. *Computational Linguistics*.

E. Bengtson and D. Roth. 2008. Understanding the value of features for coreference resolution. In *EMNLP*.

Jonathan Berant, Vivek Srikumar, Pei-Chun Chen, Abby Vander Linden, Brittany Harding, Brad Huang Peter Clark, and Christopher D. Manning. 2014. Modeling biological processes for reading comprehension. In *EMNLP*.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *ACL*.

Kai-Wei Chang, Shyam Upadhyay, Ming-Wei Chang, Vivek Srikumar, and Dan Roth. 2015. IllinoisSL: A JAVA library for structured prediction. In *Arxiv Preprint*, volume abs/1509.07179.

Agata Cybulska and Piek Vossen. 2014. Using a sledgehammer to crack a nut? lexical diversity and event coreference resolution. In *LREC*.

Agata Cybulska and Piek Vossen. 2015. Translating granularity of event slots into features for event coreference resolution. In *Workshop on Events*.

Naomi Daniel, Dragomir Radev, and Timothy Allison. 2003. Sub-event based multi-document summarization. In *the HLT-NAACL 03 on Text summarization workshop-Volume 5*.

Dipanjan Das and Noah Smith. 2011. Semi-supervised frame-semantic parsing for unknown predicates. In *ACL*.

Q. Do, Y. Chan, and D. Roth. 2011. Minimally supervised event causality identification. In *EMNLP*.

Yoav Goldberg and Michael Elhadad. 2007. SVM model tampering and anchored learning: A case study in hebrew NP chunking. In *ACL*.

Ben Hachey, Joel Nothman, and Will Radford. 2014. Cheap and easy entity evaluation. In *ACL*.

Kevin Humphreys, Robert Gaizauskas, and Saliha Azzam. 1997. Event coreference for information extraction. In *a Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*.

Heeyoung Lee, Marta Recasens, Angel Chang, Mihai Surdeanu, and Dan Jurafsky. 2012. Joint entity and event coreference resolution across documents. In *EMNLP-CoNLL*.

Zhengzhong Liu, Jun Araki, Eduard H Hovy, and Teruko Mitamura. 2014. Supervised within-document event coreference using information propagation. In *LREC*.

Xiaoqiang Luo, Sameer Pradhan, Marta Recasens, and Eduard Hovy. 2014. An extension of BLANC to system mentions.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *EMNLP*.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *ACL: System Demonstrations*.

James Mayfield, David Alexander, Bonnie J Dorr, Jason Eisner, Tamer Elsayed, Tim Finin, Clayton Fink, Marjorie Freedman, Nikesh Garera, Paul McNamee, et al. 2009. Cross-document coreference resolution: A key technology for learning by reading. In *AAAI Spring Symposium: Learning by Reading and Learning to Read*.

Srini Narayanan and Sanda Harabagiu. 2004. Question answering based on semantic structures. In *COLING*.

V. Ng and C. Cardie. 2002. Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution. In *COLING*.

Haoruo Peng, Kai-Wei Chang, and Dan Roth. 2015. A joint framework for coreference resolution and mention head detection. In *CoNLL*.

Haoruo Peng, Yangqiu Song, and Dan Roth. 2016. Event detection and co-reference with minimal supervision. In *EMNLP*, page 11. ACL.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. Conll-2011 shared task: Modeling unrestricted coreference in ontonotes. In *CoNLL*.

Sameer Pradhan, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng, and Michael Strube. 2014. Scoring coreference partitions of predicted mentions: A reference implementation. In *ACL*.

V. Punyakanok, D. Roth, and W. Yih. 2008. The importance of syntactic parsing and inference in semantic role labeling. *Computational Linguistics*, 34(2).

William M Rand. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*.

Marta Recasens and Eduard Hovy. 2011. BLANC: Implementing the rand index for coreference evaluation. *Natural Language Engineering*.

Rui Sun, Yue Zhang, Meishan Zhang, and Donghong Ji. 2015. Event-driven headline generation. In *ACL*.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *the 6th conference on Message understanding*.

Travis Wolfe, Mark Dredze, and Benjamin Van Durme. 2015. Predicate argument alignment using a global coherence model. In *NAACL*.

Bishan Yang, Claire Cardie, and Peter Frazier. 2015. A hierarchical distance-dependent bayesian model for event coreference resolution. *TACL*.

Tongtao Zhang, Hongzhi Li, Heng Ji, and Shih-Fu Chang. 2015. Cross-document event coreference resolution based on cross-media features. In *EMNLP*.