Streamlining Cross-Document Coreference Resolution: Evaluation and Modeling

Arie Cattan¹  Alon Eirew¹,²  Gabriel Stanovsky³  Mandar Joshi⁴  Ido Dagan¹
¹Bar Ilan University, Ramat-Gan, Israel  ²Intel Labs, Israel  ³The Hebrew University of Jerusalem  ⁴University of Washington, Seattle, WA
arie.cattan@gmail.com  alon.eirew@intel.com  gabis@cse.huji.ac.il  mandar90@cs.washington.edu  dagan@cs.biu.ac.il

Abstract
Recent evaluation protocols for Cross-document (CD) coreference resolution have often been inconsistent or lenient, leading to incomparable results across works and overestimation of performance. To facilitate proper future research on this task, our primary contribution is proposing a pragmatic evaluation methodology which assumes access to only raw text – rather than assuming gold mentions, disregards singleton prediction, and addresses typical targeted settings in CD coreference resolution. Aiming to set baseline results for future research that would follow our evaluation methodology, we build the first end-to-end model for this task. Our model adapts and extends recent neural models for within-document coreference resolution to address the CD coreference setting, which outperforms state-of-the-art results by a significant margin.

1 Introduction

The literature on coreference resolution has traditionally divided the task into two different settings, addressing the task at either the Within-document (WD) or Cross-document (CD) level. Each setting has presented different challenges, model design choices, and historically different evaluation practices.

In CD coreference resolution, the instances consist of multiple documents, each authored independently, without any inherent linear ordering between them. As a result, coreferring expressions across documents are often lexically-divergent, while lexically-similar expressions may refer to different concepts. Table 1 shows example documents discussing similar, yet distinct, events (two different nominations of a US Surgeon General with overlapping participants (“President Barack Obama”) and event triggers (“name”). Leveraging accurate CD coreference models seems particularly appealing for applications that merge information across texts, which have been gaining growing attention recently, such as multi-document summarization (Falke et al., 2017; Liao et al., 2018) and multi-hop question answering (Dhingra et al., 2018; Wang et al., 2019).

In this paper, we observe that research on CD coreference has been lagging behind the impressive strides made in WD coreference (Lee et al., 2017, 2018; Joshi et al., 2019, 2020), achieving 83.5 F1 (Wu et al., 2019). As the time seems ripe to promote advances in CD coreference modeling as well, we present two steps to facilitate and trigger such systematic research, with respect to proper evaluation methodologies and current modeling approaches.

With respect to evaluation, we find that previous works have often used incomparable or lenient evaluation protocols, such as assuming event and entity mentions are given as part of the input, peeking into the fine-grained subtopic annotations, or rewarding coreference models for just identifying singleton clusters (Section 2). As we will show in Section 5, these evaluation protocols have resulted in artificially inflated performance measures.

To address these shortcomings, our primary contribution consists of formalizing a more realistic evaluation methodology for CD coreference. Namely, we use only raw input texts without assuming access to human-labeled annotations such as entity and event mentions, and also disregard singletons during evaluation. In addition, we examine model performance in both focused topic clusters, known a-priory to discuss overlapping information, as well as on larger sets of documents which contain both related and unrelated documents (Section 3).

---

*Work done while at the Allen Institute for AI and the University of Washington.
Table 1: Example of sentences of topic 34 in ECB+: The underlined words represent events, same color represents a coreference cluster. Different documents describe the same event using different words (e.g. name, approached). The two subtopics present a challenging case of ambiguity in which the models need to distinguish between “name Dr Sanjay Gupta” and “name Dr. Regina Benjamin”.

With respect to modeling, in Section 4, we describe a first end-to-end CD coreference model which builds upon the state-of-the-art in WD coreference and recent advances in transformer-based encoders. To achieve this, we address the inherently non-linear nature of the CD setting by combining this model with an agglomerative clustering approach, which was shown useful in other CD models. We first show that this combination sets a new state of the art for the task of CD coreference, in comparison to prior evaluations (Section 5). We then evaluate this model following our realistic and more challenging evaluation methodology, setting a proper baseline for future research.

Taken together, our work brings the task of cross-document coreference resolution up to modern NLP standards, providing standardized evaluation benchmarks and a modern model which sets a new state-of-the-art result for the task. We hope that future work will use our framework to develop, and particularly to evaluate, models which make further advances on this challenging and important task.

2 Background: Datasets, Evaluation, and Models

We first describe the problem of within-document (WD) coreference as a reference point, in terms of established benchmark datasets, evaluation protocols, and state-of-the-art models (Section 2.1). In Section 2.2 we similarly describe the current status of cross-document (CD) coreference, which, as opposed to the WD setting, suffers from non-standard evaluation protocols and somewhat outdated models.

2.1 Within-Document Coreference

Benchmark dataset and evaluation
OntoNotes (Pradhan et al., 2012) is the standard dataset for training and testing models in the WD setting. Each document in this corpus is exhaustively annotated with both event and entity coreference clusters, while omitting singletons — entities or events which are only mentioned once and are not co-referred to in the document. The OntoNotes coreference task formulation is designed to evaluate a model’s ability to correctly identify mentions of coreferring entities and events, as well as the coreference links between them, given only the raw document text.

State-of-the-art models Models for WD coreference resolution have closely followed and adopted to recent trends in NLP, converging on end-to-end deep learning architectures which do not require intermediate structure (e.g., syntactic trees) or task-specific processing. Lee et al. (2017) presented the first prominent work to introduce recurrent neural network for the task, significantly outperforming previous works, without requiring any additional resources beside task supervision. Successive follow-up works kept improving performance through the incorporation of widely popular pretrained architectures (Lee et al., 2018), culminating recently in the introduction of the now ubiquitous BERT model (Devlin et al., 2019) to WD coreference, thus achieving the current state of the art for the task (Joshi et al., 2019; Kantor and Globerson, 2019; Wu et al., 2019).
Table 2: ECB+ statistics. # Clusters do not include singletons. The slash numbers for # Mentions, # Singletons, and # Clusters represent event/entity statistics. As recommended by the authors in the release note, we follow the split of Cybulska and Vossen (2015) that use a curated subset of the dataset.

2.2 Cross-Document Coreference

Benchmark dataset The most popular dataset for CD coreference in recent years is the ECB+ corpus (Cybulska and Vossen, 2014). Each instance in ECB+ is a set of documents, dubbed a topic, annotated with both within and cross-document coreference links (see Table 2 for details). To ensure that every instance poses challenging lexical ambiguity, each topic is a union of two sets of news documents discussing two different events (each called a subtopic), yet which are likely to use a similar vocabulary. For example, Table 1 shows two fragments from the “Obama’s announcement of Surgeon General” topic, one pertaining to the nomination of Dr. Sanjay Gupta, while the other discusses the nomination of Dr. Regina Benjamin, thus presenting a challenging disambiguation task.

As opposed to OntoNotes, only a few sentences are exhaustively annotated in each document, and the annotations include singletons. Furthermore, entities are only annotated if they participate in an event in the annotated sentence (event participants).

Evaluation The evaluation of models for CD coreference has commonly been more lenient, and less standardized than WD coreference, leading to incomparable results. This stems from three major reasons. First, while WD coreference requires models to identify entities, events, and their respective coreference links, evaluations of CD coreference mostly assumed that gold event and entity mentions are given as part of the input (Cybulska and Vossen, 2015; Kenyon-Dean et al., 2018; Barhom et al., 2019). Second, singleton clusters, as discussed in Section 2.1, are not excluded and constitute an integral part of the evaluation. Finally, CD models have been inconsistent with their usage of topic and subtopic information. Some works have evaluated performance on gold subtopics (Yang et al., 2015; Choubey and Huang, 2017), thus obviating the aforementioned designed lexical ambiguity at the topic level as criticized by Upadhyay et al. (2016). On the other hand, Barhom et al. (2019) apply the best document clustering as preprocessing, however, due to the high lexical similarity between documents within the same subtopic, this yields to an almost perfect subtopic clustering. Furthermore, evaluating only on individual subtopics disregards the fact that a coreference cluster may involve two subtopics, as we can see in Table 1 where “Barack Obama” appears in two different subtopics.

State-of-the-art models Unlike WD coreference, models for CD coreference seem to be behind the curve of recent NLP advances. The current state of the art uses intermediate structure which requires additional external resources, such as semantic role labels (Kenyon-Dean et al., 2018; Barhom et al., 2019). These models outperform simple lexical match baselines only by a few points.

3 Proposed Evaluation Methodology

In this section, we propose a standard evaluation methodology for CD coreference which addresses its main limitations, described in the previous section. First, in Section 3.1 we propose to be consistent with the WD task formulation (Pradhan et al., 2012) by: (1) assuming raw textual input without gold mention annotations; and (2) omitting singletons from the evaluation. In addition, our evaluation protocol proposes a standard breakdown of performance at topic and corpus level (Section 3.2), thus standardizing the previously non-comparable usage of topic and subtopic information.

3.1 Adapted Single-Document Standard

Raw documents as input We argue that CD coreference models should be mainly evaluated on raw text input, without assuming access to gold entity and event annotations. That is, models should perform coreference clustering of predicted rather than gold mentions. This setup is closer to the most mention identification (Yang et al., 2015; Choubey and Huang, 2017).
recent NLP task’s formulation, and while being significantly more challenging, it simulates real-world use-cases. Evaluating on gold mentions can still be valuable for error analysis, i.e., analyzing the degree to which a model erred because of incorrect mention identification vs. because of incorrect coreference linking. Finally, since only some sentences in the ECB+ annotation are (fully) annotated in each document (Section 2), the raw text evaluation should be conducted only on those sentences in order to follow the CoNLL-2012 evaluation standard.

**Omitting singletons from the evaluation** Following common practice in WD coreference, we propose to omit singleton clusters from the CD evaluation process. In fact, a model’s ability to identify that singletons do not belong to any coreference cluster is already captured in the coreference evaluation metrics. However, as shown by Rahman and Ng (2009), including singletons during the evaluation distorts the measurement of the mention-based metrics $B^3$ and CEAF$_e$ by further penalizing (or rewarding) identification of singleton span boundaries. Such penalty (or reward) is not desired when evaluating the coreference resolution task since singletons are not part of any coreference cluster. With respect to analyzes involving gold mentions, including singletons further harms the validity of the current CD evaluation protocol. Evidently, a dummy baseline which predicts no coreference links and puts each input gold mention in a singleton cluster achieves non-negligible performance (Luo, 2005) (see Tables 3 and 4).

### 3.2 Topic and Corpus Level Evaluation

As mentioned in Section 2, CD coreference models have previously made inconsistent usage of topic and subtopic information. We address this by breaking down CD model evaluation to two settings:

**Corpus level performance:** An input instance in this setting consists of a single set of documents, omitting information about the different topics and subtopics (e.g., exact number of topics) (Cybulska and Vossen, 2015; Upadhyay et al., 2016; Kenyon-Dean et al., 2018; Barhom et al., 2019). This evaluation does not make any assumption about the dataset and is also suitable for corpora in which the documents are not categorized into topics.

**Topic level performance:** Here, each gold topic is evaluated separately (Bejan and Harabagiu, 2010). In this setting, an input instance consists of a set of documents pertaining to a single topic, including, in the case of ECB+, the two subtopics which present a challenging lexical ambiguity. While this setup makes the coreference task simpler than the corpus level evaluation (Upadhyay et al., 2016), it simulates a realistic scenario where documents are initially collected at the topic level. For example, in multi-document summarization, where the goal is to generate a short summary of a topic including several documents, applying coreference resolution on the input documents has been shown to be useful for merging similar concepts (Falke et al., 2017) and generating coherent summaries (Christensen et al., 2013).

### 4 Model

Our CD model is built upon the e2e-coref single-document coreference model (Lee et al., 2017), which jointly learns mention detection and coreference link prediction, as elaborated in Section 4.1.

We modify its clustering method and training objective to port it to the cross-document setting (Section 4.2).

#### 4.1 Overview of e2e-coref

For each span $i$, the model learns a distribution $P(\cdot)$ over its possible antecedent spans $Y(i)$:

$$P(y_i) = \frac{e^{s(i,y_i)}}{\sum_{y' \in Y(i)} e^{s(i,y')}}$$

Considering all spans in a document as potential mentions, the scoring function $s(i,j)$ between span $i$ and $j$, where $j$ appears before $i$, has three components: the two mention scores $s_m(\cdot)$ of spans $i$ and $j$, and a pairwise score $s_\phi(i,j)$ for span $j$ being antecedent of span $i$.

After encoding all the tokens in a document, each possible span up to a length $n$ is represented with the concatenation of four vectors: the output representations of the span boundary (first and last) tokens, an attention-weighted sum of token representations in the span $\hat{x}_i$, and a feature vector $\phi(i)$ denoting the span length. These span representations are first fed into a mention scorer $s_m(\cdot)$ to filter the $\lambda T$ (where $T$ is the number of tokens in the document) spans with the highest scores. Then, the model learns for each of these spans to optimize the marginal log-likelihood of its correct antecedents. The full description of the different
components is described below:

\[ s(i, j) = s_m(i) + s_m(j) + s_a(i, j) \]
\[ s_m(i) = w_m \cdot \text{FFNN}_m(g_i) \]
\[ s_a(i, j) = w_a \cdot \text{FFNN}_a([g_i, g_j, g_i \circ g_j]) \]
\[ g_i = [x_{\text{FIRST}}(i), x_{\text{LAST}}(i), \hat{x}_i, \phi(i)] \]

4.2 End-to-end Cross-Document Coreference

The major obstacle in applying the e2e-coref model directly in the CD setting is its reliance on textual ordering – it forms coreference chains by linking each mention to an antecedent span appearing before it in the document. This clustering method cannot be used in the multiple-document setting since there is no inherent ordering between the documents.

Clustering Spans To overcome this challenge, we combine the model architecture from e2e-coref with an agglomerative clustering-based approach, as common in CD coreference resolution (Yang et al., 2015; Choubey and Huang, 2017; Kenyon-Dean et al., 2018; Barhom et al., 2019). The overall architecture of our model is shown in Figure 1. The agglomerative clustering step merges the most similar cluster pairs until their pairwise similarity score falls below a tuned threshold \( \tau \). Following the average-link method, the cluster pair score is defined as the average of span pair similarity scores \( s(i, j) \) (from the e2e-coref architecture) over all span pairs \( (i, j) \) across the two candidate clusters to be merged.

Training We train the model by optimizing a binary cross-entropy loss over pairs of mentions. Specifically, given a set of documents, the first step consists of encoding each document separately using RoBERTa (Liu et al., 2019). Long documents are split into non-overlapping segments of up to 512 word pieces tokens and are encoded independently (Joshi et al., 2019). For computation efficiency, we prune spans greedily, keeping only the \( \lambda T \) (where \( T \) is the total number of tokens in all the documents) highest scoring spans according to the mention scorer \( s_m(\cdot) \). Next, instead of comparing a mention only to its previous spans in the text, our pairwise scorer \( s(i, j) \) compares a mention to all other spans across all the documents.\(^4\) The positive instances for training consist of all the pairs of mentions that belong to the same coreference cluster, while the negative examples are all other pairs. To limit the computation complexity, we freeze output representations from RoBERTa instead of fine-tuning all parameters. The mention scorer \( s_m(\cdot) \) and the pairwise scorer \( s_a(i, j) \) are jointly learned to optimize the binary cross-entropy loss as follows:

\[
L = -\frac{1}{|N|} \sum_{x,z \in N} y \cdot \log(s(x, z))
\]

where \( N \) corresponds to the set of mention-pairs, and \( y \in \{0, 1\} \) to a pair label. Full implementation details are described in the appendix (Section A).

When training and evaluating the model using gold mentions, we ignore the span mention scores, \( s_m(\cdot) \), and the gold mention representations are directly fed into the pairwise scorer \( s_a(i, j) \).

Inference At inference time, we score all spans; prune spans with lowest scores; score all possible pairs; and finally form the coreference clusters using an agglomerative clustering over these pairwise scores.

Topic Level Processing To limit the search space, we apply the above algorithm separately for each topic (cluster of documents). During training, we use the gold topic segmentation of the training data. At inference time, we construct the set of topics differently for topic and corpus level evaluation (Section 3.2). We use the gold topics when

\(^4\)In practice, since the documents in ECB+ are rather short (Table 2) these pairs are mostly composed of spans from different documents.
evaluating at the topic level, as each topic is evaluated independently. However, for corpus level, since this evaluation protocol assumes the number of topics is unknown, we predict the topic clusters using another agglomerative clustering over the document representations until the document similarity drops below a threshold. Specifically, the documents are represented using TF-IDF scores of unigrams, bigrams, and trigrams, and they are merged according to their cosine similarity.

5 Empirical Assessments and Results

5.1 Empirical Assessments
This subsection assesses two aspects proposed earlier in the paper. First, as explained in Section 3.1, including singletons in the coreference evaluation is known to distort results, both in the primary evaluation over raw text (predicted mentions) as well as when evaluating performance over gold mentions as an additional analysis. Here we show empirically that in the latter setting, including singletons in the evaluation inflates results, further supporting singleton removal in such evaluation. Second, based on the same experiments, while following prior evaluation methodologies for comparison, we assess that our model substantially surpasses state-of-the-art results on ECB+, making it a suitable baseline for future research.

In order to estimate the inflation caused by the inclusion of singletons, we re-evaluated the same head-lemma baseline, the current state-of-the-art model on ECB+ (Barhom et al., 2019) and our model, while removing singletons from the predicted and gold clusters. The results are reported in Table 3 for events and in Table 4 for entities. For fair comparison, we follow Barhom et al. (2019) and cluster the documents into subtopics, using their original clustering output (Section 2).

For all models, we observe a significant performance gap when we evaluate without singletons. The differences are larger in event coreference than in entity coreference at least partly because the proportion of singletons is higher in events than in entities (30% vs 20%) in the ECB+ corpus (Table 2). The model of Barhom et al. (2019) and the same head-lemma baseline lose 11.9 and 12.3 F1 points respectively on event coreference, whereas they lose 7.4 and 9.8 F1 points on entity coreference.

The results using the MUC evaluation metric (Vilain et al., 1995) remain identical after removing singletons. This is due to the link-based characteristic of the MUC metric, which ignores singleton mentions. On the other hand, B3 and CEAFe results are much lower without singletons since both are mention-based evaluation metrics. Indeed, in addition to not being penalized for not linking these singletons into coreference clusters, models are rewarded just for predicting correctly gold mention spans.

Although removing singletons does not change the relative system ranking, it gives a better indication of the actual performance of the gold coreference resolution task, and provides a more faithful estimation of the remaining room for improvement.

Overall, our model offers an improvement of 3.5 F1 in event and 3.2 F1 in entity when ignoring singletons over the current state-of-the-art (Barhom et al., 2019) on ECB+, while surpassing the strong lemma baseline by 6.9 and 9.7 points respectively. This demonstrates that our extension of the e2e-coref accompanied with the RoBERTa encoder is a strong baseline for cross-document coreference resolution, in comparison to prior work.

5.2 Results
Here, we evaluate our model according to our proposed evaluation methodology (Section 3), in order to set the state-of-the-art baseline performance for future work on ECB+. The primary results are presented in Table 5, evaluated on predicted mentions. Additionally, Table 6 presents performance over gold mentions, allowing to analyze the impact of mention prediction. Per our methodology, in both tables results are presented for the topic and corpus levels, while ignoring singletons in the evaluation.

Since our model architecture is not tailored for a specific mention type, we use the same model separately for both event and entity coreference. In addition, inspired by Lee et al. (2012) and by the single-document standard, we encourage developing and evaluating models that perform event and entity coreference together as a single unified task, that we term “ALL”. We note that this approach is different from the JOINT model of Barhom et al. (2019), which does distinguish between event and entity mentions at test time.
larger, and (ii) this involves subtle distinction for the model (e.g. voters vs voted).

### Ambiguity

Already when evaluating on gold mentions, the performance is much lower at the topic level (Table 6) than at the subtopic level (Tables 3 and 4). Considering the nature of ECB+ where each topic consists of two similar subtopics describing two different news events (Section 2), evaluating at the subtopic level removes this designed ambiguity challenge. Indeed, the drop in precision is more significant than the drop in recall. This aspect also explains why the performance drop is more substantial in event coreference (71.1 vs. 60.8) than in entity (67.3 vs. 65.3), since the subtopics are based on event similarity. Although BERT (Devlin et al., 2019) variants have been shown to be powerful encoders in various tasks, CD coreference resolution remains challenging because of the ambiguity between coreference clusters (Table 1), even with access to gold mentions.

### Topic vs. corpus level

For all models, we see a slight performance gap between the topic and corpus level evaluation. For example, the model trained on event coreference achieves 47.6 F1 on the topic level and 46.4 on the corpus level. This demonstrates that our topic clustering algorithm, which precedes the coreference resolution step, achieves a reasonable segmentation of the documents to topics—reducing the search space without a major drop in performance. This algorithm clustered the 10 topics of the test set into 7 predicted topics. Although some gold topics were mixed, the pairwise scorer did manage to give relatively low scores to negative mention pairs across different topics.

### Predicted mentions

Overall, the performance on predicted mentions (main evaluation) is relatively low (Table 5) than that on gold mentions. This performance drop is in harmony with the single-document setting, where using gold mentions was shown to offer an improvement of 17.5 F1 (Lee et al., 2017), which corresponds to 50%
gain on error reduction. Therefore, additionally to the needed progress in making coreference decisions, there is also a large room for improvement in mention detection.

**Qualitative analysis**  We sampled topics from the development set and manually analyzed the errors of the “ALL” configuration. The most commonly occurring errors were due to an over reliance on lexical similarity. For example, the event “Maurice Cheeks was fired” was wrongly predicted to be coreferent with a similar, but distinct event, “the Sixers fired Jim O’Brien”. On the other hand, the model sometimes struggles to merge mentions which are lexically different (e.g “Jim O’Brien was shown the door”, “Jim O’Brien has been relieved”, “Philadelphia fire coach Jim O’Brien”).

The model also seems to struggle with temporal reasoning, in part, due to missing information. For example, news articles from different days have different relative reference to time, while the publication date of the articles is not always available. As a result, the model did not link “Today” in one document to “Saturday” in another document, while both referred to the same day.

### Table 5: Results of our model on the ECB+ test set using predicted mentions.

|                | MUC  | B3  | CEAF | CoNLL |
|----------------|------|-----|------|-------|
|                | R    | P   | F1   | R    | P   | F1   | R    | P   | F1   |
| Topic level    | Event| 61.1| 62   | 61.5 | 44.1| 40.9 | 42.5 | 41.6| 36.3| 38.8 | 47.6 |
|                | Entity| 40.2| 54.2| 46.2 | 23.6| 39.2 | 29.4 | 27.5| 30.2| 28.8 | 34.8 |
|                | ALL  | 49.6| 56.8| 53.0 | 32.5| 41.7 | 36.5 | 38.5| 33.5| 35.9 | 41.8 |
| Corpus level   | Event| 61.0| 61.0| 61.0 | 43.8| 38.2 | 40.8 | 39.5| 35.3| 37.3 | 46.4 |
|                | Entity| 40.8| 53.2| 46.2 | 24.1| 35.1 | 28.6 | 26.5| 28.7| 27.5 | 34.1 |
|                | ALL  | 49.3| 55.9| 52.4 | 31.5| 39.2 | 34.9 | 37.1| 32.8| 34.8 | 40.7 |

### Table 6: Results of our model on the ECB+ test set using gold mentions.

|                | MUC  | B3  | CEAF | CoNLL |
|----------------|------|-----|------|-------|
|                | R    | P   | F1   | R    | P   | F1   | R    | P   | F1   |
| Topic level    | Event| 81.0| 73.4| 77.1 | 63.3| 52.0 | 57.1 | 56.0| 42.1| 48.1 | 60.8 |
|                | Entity| 85.8| 79.3| 82.4 | 64.3| 60.0 | 62.1 | 58.6| 45.9| 51.5 | 65.3 |
|                | ALL  | 84.1| 78.2| 81.0 | 67.0| 55.1 | 60.5 | 55  | 44.5| 49.2 | 63.6 |
| Corpus level   | Event| 80.7| 72.2| 76.2 | 61.7| 48.0 | 54.0 | 53.6| 41.5| 46.8 | 59.0 |
|                | Entity| 85.3| 78.0| 81.5 | 63.5| 57.6 | 60.4 | 55.9| 43.3| 48.8 | 63.6 |
|                | ALL  | 83.8| 77.4| 80.5 | 65.8| 51.9 | 58.1 | 52.9| 46.5| 47.8 | 62.1 |

### 6 Conclusion

In this paper, we proposed a realistic evaluation methodology for cross-document coreference resolution addressing the main shortcomings of current evaluation protocols. Our proposal follows well-established standards in Within-document coreference resolution. Models are mainly evaluated on raw text while singletons are omitted during the evaluation. In addition, we formalize the usage of topic/subtopic segmentation during the evaluation for addressing the specific ambiguity challenges in CD coreference resolution.

We also established the first end-to-end baseline for CD coreference resolution, with a simple and efficient model that does not rely on external resources. Our model outperforms state-of-the-art results by 3.5 F1 points with respect to current evaluation methodologies. To the best of our knowledge, this is also the first publicly released model for cross-document coreference resolution, which is easily applicable for downstream use over raw text. Finally, we showed that when evaluating with our strict evaluation methodology, particularly when addressing the ambiguity of the corpus and topic levels (vs. sub-topics), performance dramatically decreases, suggesting a large room for improvement in future research.
References

Shany Barhom, Vered Shwartz, Alon Eirew, Michael Bugert, Nils Reimers, and Ido Dagan. 2019. Revisiting joint modeling of cross-document entity and event coreference resolution. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4179–4189, Florence, Italy. Association for Computational Linguistics.

Cosmin Bejan and Sandra Harabagiu. 2008. A linguistic resource for discovering event structures and resolving event coreference. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC’08), Marrakech, Morocco. European Language Resources Association (ELRA).

Cosmin Bejan and Sandra Harabagiu. 2010. Unsupervised event coreference resolution with rich linguistic features. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 1412–1422, Uppsala, Sweden. Association for Computational Linguistics.

Prafulla Kumar Choubey and Ruihong Huang. 2017. Event coreference resolution by iteratively unfolding inter-dependencies among events. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2124–2133, Copenhagen, Denmark. Association for Computational Linguistics.

Janara Christensen, Mausam, Stephen Soderland, and Oren Etzioni. 2013. Towards coherent multi-document summarization. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1163–1173, Atlanta, Georgia. Association for Computational Linguistics.

Agata Cybulska and Piek Vossen. 2014. Using a sledgehammer to crack a nut? lexical diversity and event coreference resolution. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14), pages 4545–4552, Reykjavik, Iceland. European Language Resources Association (ELRA).

Agata Cybulska and Piek Vossen. 2015. Translating granularity of event slots into features for event coreference resolution. In Proceedings of the The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation, pages 1–10, Denver, Colorado. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Bhuwan Dhingra, Qiao Jin, Zhihui Yang, William Cohen, and Ruslan Salakhutdinov. 2018. Neural models for reasoning over multiple mentions using coreference. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 42–48, New Orleans, Louisiana. Association for Computational Linguistics.

Tobias Falke, Christian M. Meyer, and Iryna Gurevych. 2017. Concept-map-based multi-document summarization using concept coreference resolution and global importance optimization. In Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 801–811, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the thirteenth international conference on artificial intelligence and statistics, pages 249–256.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. Transactions of the Association for Computational Linguistics, 8:64–77.

Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel Weld. 2019. BERT for coreference resolution: Baselines and analysis. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 5803–5808, Hong Kong, China. Association for Computational Linguistics.

Ben Kantor and Amir Globerson. 2019. Coreference resolution with entity equalization. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 673–677, Florence, Italy. Association for Computational Linguistics.

Kian Kenyon-Dean, Jackie Chi Kit Cheung, and Doina Precup. 2018. Resolving event coreference with supervised representation learning and clustering-oriented regularization. In Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, pages 1–10, New Orleans, Louisiana. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Heeyoung Lee, Marta Recasens, Angel Chang, Mihai Surdeanu, and Dan Jurafsky. 2012. Joint entity and
event coreference resolution across documents. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 489–500, Jeju Island, Korea. Association for Computational Linguistics.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

Kenton Lee, Luheng He, and Luke S. Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In Proceedings of the 2018 Annual Conference of the North American Chapter of the Association for Computational Linguistics.

Kexin Liao, Logan Lebanoff, and Fei Liu. 2018. Abstract Meaning Representation for multi-document summarization. In Proceedings of the 27th International Conference on Computational Linguistics, pages 1178–1190, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, pages 25–32, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In Joint Conference on EMNLP and CoNLL - Shared Task, pages 1–40, Jeju Island, Korea. Association for Computational Linguistics.

Altaf Rahman and Vincent Ng. 2009. Supervised models for coreference resolution. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 968–977, Singapore. Association for Computational Linguistics.

Shyam Upadhyay, Nitish Gupta, Christos Christodoulopoulos, and Dan Roth. 2016. Revisiting the evaluation for cross document event coreference. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 1949–1958, Osaka, Japan. The COLING 2016 Organizing Committee.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995.

Haoyu Wang, Mo Yu, Xiaoxiao Guo, Rajarshi Das, Wenhan Xiong, and Tian Gao. 2019. Do multi-hop readers dream of reasoning chains? In Proceedings of the 2nd Workshop on Machine Reading for Question Answering, pages 91–97, Hong Kong, China. Association for Computational Linguistics.

Wei Wu, Fei Wang, Arianna Yuan, Fei Wu, and Jiwei Li. 2019. Coreference resolution as query-based span prediction. arXiv preprint arXiv:1911.01746.

Bishan Yang, Claire Cardie, and Peter Frazier. 2015. A hierarchical distance-dependent Bayesian model for event coreference resolution. Transactions of the Association for Computational Linguistics, 3:517–528.

A Implementation Details

Learning Our model includes 14M parameters and is implemented in PyTorch (Paszke et al., 2017), using the Adam optimizer (Kingma and Ba, 2014) with a batch size of 32. The layers of the models are initialized with Xavier Glorot method (Glorot and Bengio, 2010). The span pruning coefficient $\lambda$ is set to 0.25 for event, 0.35 for entity and 0.45 for ALL. The maximum span length $n$ is 10 for events and 15 for entity and ALL. For all experiments, we use a single 12GB GPU, where the training takes a few hours in the worst case (ALL on predicted mentions) and the inference only a few minutes.

Hyperparameters For each model, the agglomerative clustering threshold $\tau$ was manually tuned on the development set in range {0.5, 0.65} to maximize the CoNLL F1 score. For models trained on gold mentions, $\tau$ is set to 0.6, 0.6, and 0.65 for event, entity and ALL respectively. For models trained on raw text, $\tau$ is set to 0.65, 0.65, and 0.55 for event, entity and ALL respectively.