

# CDLM: Cross-Document Language Modeling

Avi Caciularu<sup>1\*</sup> Arman Cohan<sup>2,3</sup> Iz Beltagy<sup>2</sup>  
Matthew E. Peters<sup>2</sup> Arie Cattan<sup>1</sup> Ido Dagan<sup>1</sup>

<sup>1</sup>Computer Science Department, Bar-Ilan University, Ramat-Gan, Israel

<sup>2</sup>Allen Institute for Artificial Intelligence, Seattle, WA

<sup>3</sup>Paul G. Allen School of Computer Science & Engineering, University of Washington

avi.c33@gmail.com, {armanc,beltagy,matthewp}@allenai.org

arie.cattan@gmail.com, dagan@cs.biu.ac.il

## Abstract

We introduce a new pretraining approach geared for multi-document language modeling, incorporating two key ideas into the masked language modeling self-supervised objective. First, instead of considering documents in isolation, we pretrain over sets of multiple related documents, encouraging the model to learn cross-document relationships. Second, we improve over recent long-range transformers by introducing dynamic global attention that has access to the entire input to predict masked tokens. We release CDLM (Cross-Document Language Model), a new general language model for multi-document setting that can be easily applied to downstream tasks. Our extensive analysis shows that both ideas are essential for the success of CDLM, and work in synergy to set new state-of-the-art results for several multi-text tasks.<sup>1</sup>

## 1 Introduction

The majority of NLP research addresses a *single* text, typically at the sentence or document level. Yet, there are important applications which are concerned with aggregated information spread across multiple texts, such as cross-document coreference resolution (Cybulska and Vossen, 2014), classifying relations between document pairs (Zhou et al., 2020) and multi-hop question answering (Yang et al., 2018).

Existing language models (LMs) (Devlin et al., 2019a; Liu et al., 2019; Raffel et al., 2020), which are pretrained with variants of the masked language modeling (MLM) self-supervised objective, are known to provide powerful representations for internal text structure (Clark et al., 2019; Rogers et al., 2020a), which were shown to be beneficial

Doc 1: "Harry Shearer is **suing** Vivendi's Universal Music for \$125 million for allegedly fraudulent ..."

Doc 2: "...Harry Shearer **alleges** parent company of Universal Music and StudioCanal withheld millions..."

Doc 3: "Shearer was then **joined in the lawsuit** with StudioCanal and its French parent Vivendi by his co-stars"

Figure 1: An example from Multi-News (Fabbri et al., 2019). Circled words represent matching events and the same color represents mention alignments.

also for various multi-document tasks (Yang et al., 2020; Zhou et al., 2020).

In this paper, we point out that beyond modeling internal text structure, multi-document tasks require also modeling cross-text relationships, particularly aligning or linking matching information elements across documents. For example, in Fig. 1, one would expect a competent model to correctly capture that the two event mentions *suing* and *alleges*, from Documents 1 and 2, should be matched. Accordingly, capturing such cross-text relationships, in addition to representing internal text structure, can prove useful for downstream multi-text tasks, as we demonstrate empirically later.

Following this intuition, we propose a new simple cross-document pretraining procedure, which is applied over sets of *related* documents, in which informative cross-text relationships are abundant (e.g. like those in Fig. 1). Under this setting, the model is encouraged to learn to consider and represent such relationships, since they provide useful signals when optimizing for the language modeling objective. For example, we may expect that it will be easier for a model to unmask the word *alleges* in Document 2 if it would manage to effectively “peek” at Document 1, by matching the masked position and its context with the corresponding information in the other document.

Naturally, considering cross-document context in pretraining, as well as in finetuning, requires

\* Work partly done as an intern at AI2.

<sup>1</sup>Code and models are available at <https://github.com/aviclu/CDLM>

a model that can process a fairly large amount of text. To that end, we leverage recent advances in developing efficient long-range transformers (Beltagy et al., 2020; Zaheer et al., 2020), which utilize a global attention mode to build representations based on the entire input. Overcoming certain restrictions in prior utilization of global attention (see Section 2.1), we introduce a dynamic attention pattern during pretraining, over all masked tokens, and later utilize it selectively in finetuning.

Combining pretraining over related documents along with our global attention pattern yields a novel pretraining approach, that is geared to learn and implicitly encode informative cross-document relationships. As our experiments demonstrate, the resulting model, termed Cross-Document Language Model (CDLM), can be generically applied to downstream multi-document tasks, eliminating the need for task-specific architectures. We show empirically that our model improves consistently over previous approaches in several tasks, including cross-document coreference resolution, multi-hop question answering, and document matching tasks. Moreover, we provide controlled experiments to ablate the two contributions of pretraining over related documents as well as new dynamic global attention. Finally, we provide additional analyses that shed light on the advantageous behavior of our CDLM. Our contributions are summarized below:

- A new pretraining approach for multi-document tasks utilizing: (1) sets of related documents instead of single documents; (2) a new dynamic global attention pattern.
- The resulting model advances the state-of-the-art for several multi-document tasks.

## 2 Method

### 2.1 Background: the Longformer Model

Recently, long-range LMs (e.g., Longformer (Beltagy et al., 2020), BigBird (Zaheer et al., 2020)) have been proposed to extend the capabilities of earlier transformers (Vaswani et al., 2017) to process long sequences, using a sparse self-attention architecture. These models showed improved performance on both long-document and multi-document tasks (Tay et al., 2021). In the case of multiple documents, instead of encoding documents separately, these models allow concatenating them into a long sequence of tokens and encoding them jointly. We base our model on Longformer, which sparsifies

the full self-attention matrix in transformers by using a combination of a localized sliding window (called local attention), as well as a global attention pattern on a few specific input locations. Separate weights are used for global and local attention. During pretraining, Longformer assigns *local attention* to all tokens in a window around each token and optimizes the Masked Language Modeling (MLM) objective. Before task-specific finetuning, the attention mode is predetermined for each input token, assigning global attention to a few targeted tokens, such as special tokens, that are targeted to encode global information. Thus, in the Longformer model, global attention weights are not pretrained. Instead, they are initialized to the local attention values, before finetuning on each downstream task. We conjecture that the global attention mechanism can be useful for learning meaningful representations for modeling cross-document (CD) relationships. Accordingly, we propose augmenting the pretraining phase to exploit the global attention mode, rather than using it only for task-specific finetuning, as described below.

### 2.2 Cross-Document Language Modeling

We propose a new pretraining approach consisting of two key ideas: (1) pretraining over sets of *related* documents that contain overlapping information (2) pretraining with a dynamic global attention pattern over masked tokens, for referencing the entire cross-text context.

**Pretraining Over Related Documents** Documents that describe the same topic, e.g., different news articles discussing the same story, usually contain overlapping information. Accordingly, various CD tasks may leverage from an LM infrastructure that encodes information regarding alignment and mapping across multiple texts. For example, for the case of CD coreference resolution, consider the underlined predicate examples in Figure 1. One would expect a model to correctly align the mentions denoted by *suing* and *alleges*, effectively recognizing their cross-document relation.

Our approach to cross-document language modeling is based on pretraining the model on sets (clusters) of documents, all describing the same topic. Such document clusters are readily available in a variety of existing CD benchmarks, such as multi-document summarization (e.g., Multi-News (Fabbri et al., 2019)) and CD coreference resolution (e.g., ECB+ (Cybulska and Vossen, 2014)). Pre-

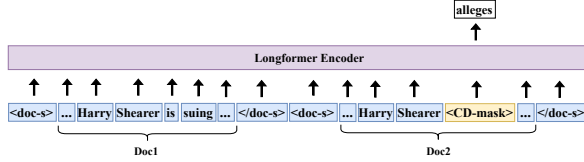


Figure 2: CDLM pretraining: The input consists of concatenated documents, separated by special document separator tokens. The masked (unmasked) token colored in yellow (blue) represents global (local) attention. The goal is to predict the masked token *alleges*, based on the global context, i.e., the entire set of documents.

training the model over a set of related documents encourages the model to learn cross-text mapping and alignment capabilities, which can be leveraged for improved unmasking, as exemplified in Sec. 1. Indeed, we show that this strategy directs the model to utilize information across documents and helps in multiple downstream CD tasks.

**Pretraining With Global Attention** To support contextualizing information across multiple documents, we need to use efficient transformer models that scale linearly with input length. Thus, we base our cross-document language model (CDLM) on the Longformer model (Beltagy et al., 2020), however, our setup is general and can be applied to other similar efficient Transformers. As described in Sec. 2.1, Longformer sparsifies the expensive attention operation for long inputs using a combination of local and global attention modes. As input to the model, we simply concatenate related documents using new special document separator tokens,  $\langle \text{doc-s} \rangle$  and  $\langle / \text{doc-s} \rangle$ , for marking document boundaries. We apply a similar masking procedure as in BERT: For each training example, we randomly choose a sample of tokens (15%) to be masked;<sup>2</sup> however, our pretraining strategy tries to predict each masked token while considering the *full* document set, by assigning them *global attention*, utilizing the global attention weights (see Section 2.1). This allows the Longformer to contextualize information both across documents as well as over long-range dependencies within-document. The non-masked tokens use local attention, by utilizing the local attention weights, as usual.

An illustration of the CD pretraining procedure is depicted in Fig. 2, where the masked token associated with *alleges* (colored in yellow) globally attends to the whole sequence, and the rest of the non-masked tokens (colored in blue) attend to their local context. With regard to the example in Fig. 1,

<sup>2</sup>For details of masking see BERT (Devlin et al., 2019b).

this masking approach aims to implicitly compel the model to learn to correctly predict the word *alleges* by looking at the second document, optimally at the phrase *suing*, and thus capture the alignment between these two events and their contexts.

### 2.3 CDLM Implementation

In this section, we provide the experimental details used for pretraining our CDLM model.

**Corpus data** We use the preprocessed Multi-News dataset (Fabbri et al., 2019) as the source of related documents for pretraining. This dataset contains 44,972 training document clusters, originally intended for multi-document summarization. The number of source documents (that describe the same topic) per cluster varies from 2 to 10, as detailed in Appendix A.1. We consider each cluster of at least 3 documents for our cross-document pretraining procedure. We compiled our training corpus by concatenating related documents that were sampled randomly from each cluster, until reaching the Longformer’s input sequence length limit of 4,096 tokens per sample. Note that this pretraining dataset is relatively small compared to conventional datasets used for pretraining. However, using it results in the powerful CDLM model.

**Training and hyperparameters** We pretrain the model according to our pretraining strategy, described in Section 2.2. We employ the Longformer-base model (Beltagy et al., 2020) using the HuggingFace implementation (Wolf et al., 2020) and continue its pretraining, over our training data, for an additional 25k steps.<sup>3</sup> The new document separator tokens are added to the model vocabulary and randomly initialized before pretraining. We use the same setting and hyperparameters as in Beltagy et al. (2020), and as elaborated in Appendix B.

## 3 Evaluations and Results

This section presents experiments conducted to evaluate our CDLM, as well as the ablations and baselines we used. For the intrinsic evaluation we measured the perplexity of the models. For extrinsic evaluations we considered event and entity cross-document coreference resolution, paper citation recommendation, document plagiarism detection, and multihop question answering. We also

<sup>3</sup>The training process for the base model takes 8 days on 8 RTX8000 GPUs. Training large models requires roughly 3x compute; therefore we do not focus on large models here and leave that for future work.

conducted an attention analysis, showing that our CDLM indeed captured cross-document and long-range relations during pretraining.<sup>4</sup>

**Baseline LMs** Recall that CDLM employs multiple related documents during pretraining, and assigns global attention to masked tokens. To systematically study the importance of these two components, we consider the following LM baselines:

- **Longformer**: the underlying Longformer model, without additional pretraining.
- **Local CDLM**: pretrained using the same corpus of CDLM with the Longformer’s attention pattern (local attention only). This baseline is intended to separate the effect of simply continuing pretraining Longformer on our new pre-training data.
- **Rand CDLM**: Longformer with the additional CDLM pretraining, while using random, unrelated documents from various clusters. This baseline model allows assessing whether pretraining using related documents is beneficial.
- **Prefix CDLM**: pretrained similarly as CDLM but uses global attention for the first tokens in the input sequence, rather than the masked ones. This resembles the attention pattern of BIGBIRD (Zaheer et al., 2020), adopted for our cross-document setup. We use this ablation for examining this alternative global attention pattern, from prior work.

The data and pretraining hyperparameters used for the ablations above are the same as the ones used for our CDLM pretraining, except for the underlying Longformer, which is not further pre-trained, and the Rand CDLM, that is fed with different document clusters (drawn from the same corpus). During all the experiments, the global attention weights used by the underlying Longformer and by Local CDLM are initialized to the values of their pretrained local attention weights. All the models above further finetune their global attention weights, depending on the downstream task. When finetuning CDLM and the above models on downstream tasks involving multiple documents, we truncate the longer inputs to the Longformer’s 4,096 token limit.

### 3.1 Cross-Document Perplexity

First, we conduct a cross-document (CD) perplexity experiment, in a task-independent manner, to as-

<sup>4</sup>Since the underlying Longformer model is encoder-only, we evaluate on tasks that can be modeled using the encoder-only setting. We leave extensions to address seq2seq tasks like generation to future work.

Model	Validation	Test
Longformer	3.89	3.94
Local CDLM	3.78	3.84
Rand CDLM	3.68	3.81
Prefix CDLM	<b>3.20</b>	3.41
CDLM	3.23	<b>3.39</b>

Table 1: Cross-document perplexity evaluation on the validation and tests set of Multi-News. Lower is better.

sess the contribution of the pretraining process. We used the Multi-News validation and test sets, each of them containing 5,622 document clusters, to construct the evaluation corpora. Then we followed the same protocol from the pretraining phase - 15% of the input tokens are randomly masked, where the challenge is to predict the masked token given all documents in the input sequence. We matched the pretraining phase of each one of the ablation models: In CDLM and Rand CDLM, we assigned global attention for the masked tokens, and for Prefix CDLM the global attention is assigned to the 15% first input tokens. Both Longformer and Local CDLM used local attention only. Perplexity is then measured by computing exponentiation of the loss.

The results are depicted in Table 1. The advantage of CDLM over Rand CDLM, which was pretrained equivalently over an equivalent amount of (unrelated) CD data, confirms that CD pretraining, over *related* documents, indeed helps for CD masked token prediction across such documents. Prefix CDLM introduces similar results since it was pretrained using a global attention pattern and the same corpora used by CDLM. The Local CDLM is expected to have difficulty to predict tokens across documents since it was pretrained without using global attention. Finally, the underlying Longformer model, which is reported as a reference point, is inferior to all the ablations since it was pretrained in a single document setting and without global attention or further pretraining on this domain. Unlike the two local-attentive models, CDLM is encouraged to look at the full sequence when predicting a masked token. Therefore, as in the pretraining phase, it exploits related information in other documents, and not just the local context of the masked token, hence CDLM, as well as Prefix CDLM, result with a substantial performance gain.

### 3.2 Cross-Document Coreference Resolution

Cross-document (CD) coreference resolution deals with identifying and clustering together textual



mentions across multiple documents that refer to the same concept (see Fig. 1). The considered mentions can be either entity mentions, usually noun phrases, or event mentions, typically verbs or nominalizations that appear in the text.

**Benchmark.** We evaluated our CDLM by utilizing it over the ECB+ corpus (Cybulska and Vossen, 2014), the most commonly used dataset for CD coreference. ECB+ consists of within- and cross-document coreference annotations for entities and events (statistics are given in Appendix A.2). Following previous work, for comparison, we conduct our experiments on gold event and entity mentions.

We follow the standard coreference resolution evaluation metrics: *MUC* (Vilain et al., 1995), *B<sup>3</sup>* (Bagga and Baldwin, 1998), *CEAF<sub>e</sub>* (Luo, 2005), their average *CoNLL F1*, and the more recent *LEA* metric (Moosavi and Strube, 2016).

**Algorithm.** Recent approaches for CD coreference resolution train a pairwise scorer to learn the probability that two mentions are co-referring. At inference time, an agglomerative clustering based on the pairwise scores is applied, to form the coreference clusters. We made several modifications to the pairwise scorer. The current state-of-the-art models (Zeng et al., 2020; Yu et al., 2020) train the pairwise scorer by including only the local contexts (containing sentences) of the candidate mentions. They concatenate the two input sentences and feed them into a transformer-based LM. Then, part of the resulting tokens representations are aggregated into a single feature vector which is passed into an additional MLP-based scorer to produce the coreference probability estimate. To accommodate our proposed CDLM model, we modify this modeling by including the entire documents containing the two candidate mentions, instead of just their containing sentences, and assigning the global attention mode to the mentions’ tokens and to the [CLS] token. The full method and hyperparameters are elaborated in Appendix C.1.

**Baselines.** We consider state-of-the-art baselines that reported results over the ECB+ benchmark. The following baselines were used for both event and entity coreference resolution:

- Barhom et al. (2019) is a model trained jointly for solving event and entity coreference as a single task. It utilizes semantic role information between the candidate mentions.

- Cattan et al. (2020) is a model trained in an end-to-end manner (jointly learning mention detection and coreference following Lee et al. (2017)), employing the RoBERTa-large model to encode each document separately and to train a pair-wise scorer atop.

- Allaway et al. (2021) is a BERT-based model combining sequential prediction with incremental clustering.

The following baselines were used for event coreference resolution. They all integrate external linguistic information as additional features.

- Meged et al. (2020) is an extension of Barhom et al. (2019), leveraging external knowledge acquired from a paraphrase resource (Shwartz et al., 2017).

- Zeng et al. (2020) is an end-to-end model, encoding the concatenated two sentences containing the two mentions by the BERT-large model. Similarly to our algorithm, they feed a MLP-based pairwise scorer with the concatenation of the [CLS] representation and an attentive function of the candidate mentions representations.

- Yu et al. (2020) is an end-to-end model similar to Zeng et al. (2020), but uses rather RoBERTa-large and does not consider the [CLS] contextualized token representation for the pairwise classification.

**Results.** The results on event and entity CD coreference resolution are depicted in Table 2. Our CDLM outperforms all methods, including the recent sentence based models on event coreference. All the results are statistically significant using bootstrap and permutation tests with  $p < 0.001$  (Dror et al., 2018). CDLM largely surpasses state-of-the-art results on entity coreference, even though these models utilize external information and use large pretrained models, unlike our base model. In Table 3, we provide the ablation study results. Using our model with sentences only, i.e., considering only the sentences where the candidate mentions appear (as the prior baselines did), exhibits lower performance, resembling the best performing baselines. Some crucial information about mentions can appear in a variety of locations in the document, and is not concentrated in one sentence. This characterizes long documents, where pieces of information are often spread out. Overall, the ablation study shows the advantage of using our pretraining method, over related documents and using a scattered global attention pattern, com-

		MUC			$B^3$			$CEAF_e$			LEA			CoNLL
		R	P	$F_1$	R	P	$F_1$	R	P	$F_1$	R	P	$F_1$	$F_1$
Event	Barhom et al. (2019)	78.1	84.0	80.9	76.8	86.1	81.2	79.6	73.3	76.3	64.6	72.3	68.3	79.5
	Meged et al. (2020)	78.8	84.7	81.6	75.9	85.9	80.6	81.1	74.8	77.8	64.7	73.4	68.8	80.0
	Cattan et al. (2020)	85.1	81.9	83.5	82.1	82.7	82.4	75.2	78.9	77.0	68.8	72.0	70.4	81.0
	Zeng et al. (2020)	85.6	89.3	87.5	77.6	89.7	83.2	84.5	80.1	<b>82.3</b>	-	-	-	84.3
	Yu et al. (2020)	88.1	85.1	86.6	86.1	84.7	85.4	79.6	83.1	81.3	-	-	-	84.4
	Allaway et al. (2021)	81.7	82.8	82.2	80.8	81.5	81.1	79.8	78.4	79.1	-	-	-	80.8
	CDLM	87.1	89.2	<b>88.1</b>	84.9	87.9	<b>86.4</b>	83.3	81.2	82.2	76.7	77.2	<b>76.9</b>	<b>85.6</b>
Entity	Barhom et al. (2019)	81.0	80.8	80.9	66.8	75.5	70.9	62.5	62.8	62.7	53.5	63.8	58.2	71.5
	Cattan et al. (2020)	85.7	81.7	83.6	70.7	74.8	72.7	59.3	67.4	63.1	56.8	65.8	61.0	73.1
	Allaway et al. (2021)	83.9	84.7	84.3	74.5	70.5	72.4	70.0	68.1	69.2	-	-	-	75.3
	CDLM	88.1	91.8	<b>89.9</b>	82.5	81.7	<b>82.1</b>	81.2	72.9	<b>76.8</b>	76.4	73.0	<b>74.7</b>	<b>82.9</b>

Table 2: Results on event and entity cross-document coreference resolution on ECB+ test set.

	F1	$\Delta$
full document CDLM	85.6	
– sentences only CDLM	84.2	-1.4
– Longformer	84.6	-1.0
– Local CDLM	84.7	-0.9
– Rand CDLM	84.1	-1.5
– Prefix CDLM	85.1	-0.5

Table 3: Ablation results (CoNLL F1) on our model on the test set of ECB+ event coreference.

pared to the other examined settings. Recently, our CDLM-based coreference model was utilized to generate event clusters within an effective faceted-summarization system for multi-document exploration (Hirsch et al., 2021).

### 3.3 Document matching

We evaluate our CDLM over document matching tasks, aiming to assess how well our model can capture interactions across multiple documents. We use the recent multi-document classification benchmark by Zhou et al. (2020) which includes two tasks of citation recommendation and plagiarism detection. The goal of both tasks is categorizing whether a particular relationship holds between two input documents. Citation recommendation deals with detecting whether one reference document should cite the other one, while the plagiarism detection task infers whether one document plagiarizes the other one. To compare with recent state-of-the-art models, we utilized the setup and data selection from Zhou et al. (2020), which provides three datasets for citation recommendation and one for plagiarism detection.

**Benchmarks.** For citation recommendation, the datasets include the ACL Anthology Network Corpus (AAN; Radev et al., 2013), the Semantic Scholar Open Corpus (OC; Bhagavatula et al.,

2018), and the Semantic Scholar Open Research Corpus (S2ORC; Lo et al., 2020). For plagiarism detection, the dataset is the Plagiarism Detection Challenge (PAN; Potthast et al., 2013).

AAN is composed of computational linguistics papers which were published on the ACL Anthology from 2001 to 2014, OC is composed of computer science and neuroscience papers, S2ORC is composed of open access papers across broad domains of science, and PAN is composed of web documents that contain several kinds of plagiarism phenomena. For further dataset preprocessing details and statistics, see Appendix A.3.

**Algorithm.** For our models, we added the [CLS] token at the beginning of the input sequence, assigned it global attention, and concatenated the pair of texts, according to the finetuning setup discussed in Section 2.2. The hyperparameters are further detailed in Appendix C.2.

**Baselines.** We consider the reported results of the following recent baselines:

- **HAN** (Yang et al., 2016) proposed the Hierarchical Attention Networks (HANs). These models employ a bottom-up approach in which a document is represented as an aggregation of smaller components i.e., sentences, and words. They set competitive performance in different tasks involving long document encoding (Sun et al., 2018).
- **SMASH** (Jiang et al., 2019) is an attentive hierarchical recurrent neural network (RNN) model, used for tasks related to long documents.
- **SMITH** (Yang et al., 2020) is a BERT-based hierarchical model, similar HANs.
- **CDA** (Zhou et al., 2020) is a cross-document attentive mechanism (CDA) built on top of HANs, based on BERT or GRU models (see Section 4).

Model	AAN	OC	S2orc	PAN
SMASH (2019) <sup>5</sup>	80.8	-	-	-
SMITH (2020) <sup>5</sup>	85.4	-	-	-
BERT-HAN (2020)	65.0	86.3	90.8	<b>87.4</b>
GRU-HAN+CDA (2020)	75.1	89.9	91.6	78.2
BERT-HAN+CDA (2020)	82.1	87.8	92.1	86.2
Longformer	85.4	93.4	95.8	80.4
Local CDLM	83.8	92.1	94.5	80.9
Rand CDLM	85.7	93.5	94.6	79.4
Prefix CDLM	87.3	94.8	94.7	81.7
CDLM	<b>88.8</b>	<b>95.3</b>	<b>96.5</b>	82.9

Table 4:  $F_1$  scores over the document matching benchmarks’ test sets.

Both SMASH and SMITH reported results only over the AAN benchmark. In addition, they used a slightly different version of the AAN dataset,<sup>5</sup> and included the full documents, unlike the dataset that (Zhou et al., 2020) used, which we utilized as well, that considers only the documents’ abstracts.

**Results.** The results on the citation recommendation and plagiarism detection tasks are depicted in Table 4. We observe that even though SMASH and SMITH reported results using the full documents for the AAN task, our model outperforms them, using the partial version of the dataset, as in Zhou et al. (2020). Moreover, unlike our model, CDA is task-specific since it trains new cross-document weights for each task, yet it is still inferior to our model, evaluating on the three citation recommendation benchmarks. On the plagiarism detection benchmark, interestingly, our models does not perform better. Moreover, CDA impairs the performance of BERT-HAN, implying that dataset does not require detailed cross-document attention at all. In our experiments, finetuning BERT-HAN+CDA over the PAN dataset yielded poor results:  $F_1$  score of 79.6, substantially lower compared to our models. The relatively small size of PAN may explain such degradations.

### 3.4 Multihop Question answering

In the task of multihop question answering, a model is queried to extract answer spans and evidence sentences, given a question and multiple paragraphs from various related and non-related documents. This task includes challenging questions, that answering them requires finding and reasoning over

<sup>5</sup>Following the most recent work of Zhou et al. (2020), we evaluate our model on their version of the dataset. We also quote the results of SMASH and SMITH methods, even though they used a somewhat different version of this dataset, hence their results are not fully comparable to the results of our model and those of CDA.

Model	Ans	Sup	Joint
Transformer-XH (2020)	66.2	72.1	52.9
Graph Recurrent Retriever (2020)	73.3	76.1	61.4
RoBERTa-lf (2020)	73.5	83.4	63.5
BIGBIRD (2020)	<b>75.5</b>	<b>87.1</b>	<b>67.8</b>
Longformer	74.5	83.9	64.5
Local CDLM	74.1	84.0	64.2
Rand CDLM	72.7	84.8	63.7
Prefix CDLM	74.8	84.7	65.2
CDLM	74.7	86.3	66.3

Table 5: HotpotQA-distractor results ( $F_1$ ) for the dev set. We use the “base” model size results from prior work for direct comparison. Ans: answer span, Sup: Supporting facts.

multiple supporting documents.

**Benchmark.** We used the HotpotQA-distractor dataset (Yang et al., 2018). Each example in the dataset is comprised of a question and 10 different paragraphs from different documents, extracted from Wikipedia; two gold paragraphs include the relevant information for properly answering the question, mixed and shuffled with eight distractor paragraphs (for the full dataset statistics, see Yang et al. (2018)). There are two goals for this task: extraction of the correct answer span, and detecting the supporting facts, i.e., evidence sentences.

**Algorithm.** We employ the exact same setup from (Beltagy et al., 2020): We concatenate all the 10 paragraphs into one large sequence, separated by document separator tokens, and using special sentence tokens to separate sentences. The model is trained jointly in a multi-task manner, where classification heads specialize on each sub-task, including relevant paragraphs prediction, evidence sentences identification, extracting answer spans and inferring the question types (yes, no, or span). For details and hyperparameters, see Appendix C.3 and Beltagy et al. (2020, Appendix D).

**Results.** The results are depicted in Table 5, where we included also the results for Transformer-XH (Zhao et al., 2020), a transformer-based model that constructs global contextualized representations, Graph Recurrent Retriever (Asai et al., 2020), a recent strong graph-based passage retrieval method, RoBERTa (Liu et al., 2019), which was modified by Beltagy et al. (2020) to operate on long sequences (dubbed RoBERTa-lf), and BIGBIRD (Zaheer et al., 2020), a long-range transformer model which was pretrained on a massive amount of text. CDLM outperforms all the ablated models as well as the comparably sized models

from prior work (except for BIGBIRD), especially in the supporting evidence detection sub-task. We note that the BIGBIRD model was pretrained on much larger data, using more compute resources compared both to the Longformer model and to our models. We suspect that with more compute and data, it is possible to close the gap between CDLM and BIGBIRD performance. We leave for future work evaluating a larger version of the CDLM model against large, state-of-the-art models.

### 3.5 Attention Analysis

It was recently shown that during the pretraining phase, LMs learn to encode various types of linguistic information, that can be identified via their attention patterns (Wiegrefe and Pinter, 2019; Rogers et al., 2020b). In Clark et al. (2019), the attention weights of BERT were proved as informative for probing the degree to which a particular token is “important”, as well as its linguistic roles. For example, they showed that the averaged attention weights from the last layer of BERT are beneficial features for dependency parsing.

We posit that our pretraining scheme, which combines global attention and a multi-document context, captures alignment and mapping information across documents. Hence, we hypothesize that the global attention mechanism favors cross-document (CD), long-range relations. To gain more insight, our goal is to investigate if our proposed pretraining method leads to relatively higher global attention weights between co-referring mentions compared to non-co-referring ones, even without any finetuning over CD coreference resolution.

**Benchmark.** We randomly sampled 1,000 positive and 1,000 negative coreference-pair examples from the ECB+ CD coreference resolution benchmark, for both events and entities. Each example consists of two concatenated documents and two coreference candidate mentions (see Section 3.2).

**Analysis Method.** For each example, which contains two mention spans, we randomly pick one to be considered as the *source span*, while the second one is the *target span*. We denote the set of the tokens in the source and target spans as  $S$  and  $T$ , respectively. Our goal is to quantify the degree of alignment between  $S$  and  $T$ , using the attention pattern of the model. We first assign global attention to the tokens in the source span (in  $S$ ). Next, we pass the full input through the model, compute the normalized attention weights for all the tokens

---

**Doc 1:** President Obama will name Dr. Regina Benjamin as U.S. Surgeon General in a Rose Garden announcement late this morning. Benjamin, an Alabama family physician, [...]

**Doc 2:** [...] Obama nominates new surgeon general: MacArthur “genius grant” fellow Regina Benjamin. [...]

---

Figure 3: An example from ECB+ corpus. The underlined phrases represent a positive, co-referring event mention pair. The blue (green) colored mention is considered as the *source* (*target*) span.

in the input with respect to  $S$ , by aggregating the scores extracted from the last layer of the model. The score for an input token  $i \notin S$ , is given by

$$s(i|S) \propto \exp \left[ \sum_{k=1}^n \sum_{j \in S} \left( \alpha_{i,j}^k + \alpha_{j,i}^k \right) \right],$$

where  $\alpha_{i,j}^k$  is the *global* attention weight from token  $i$  to token  $j$  produced by head  $k$ , and  $n$  is the total number of attention heads (the score is computed using only the last layer of the model). Note that we include both directions of attention. The target span score is then given by  $s(T|S) = \frac{1}{|T|} \sum_{j \in T} s(j|S)$ . Finally, we calculate the percentile rank (PR) of  $s(T|S)$ , compared to the rest of the token scores within the containing document of  $T$ , namely,  $\{s(i|S) | i \notin T\}$ .

For positive coreference examples, plausible results are expected to be associated with high attention weights between the source and the target spans, resulting with a high value of  $s(T|S)$ , and thus, yielding a higher PR. For negative examples, the target span is not expected to be promoted with respect to the rest of the tokens in the document.

**Results.** First, we apply the procedure above over one selected example, depicted in Figure 3. We consider the two CD co-referring event mentions: *name* and *nominates* as the source and target spans, respectively. The target span received a PR of 69% when evaluating the underlying Longformer. Notably, it received a high PR of 90% when using our CDLM, demonstrating the advantage of our novel pretraining method. Next, we turn to a systematic experiment, elucidating the relative advantage of pretraining with global attention across related documents. In Table 6, we depict the mean PR (MPR) computed over all the sampled examples, for all our pretrained models. We observe that none of the models fail<sup>6</sup> on the set of negatives, since the negative examples contain reasonable event or entity mentions, rather than random, non informative

<sup>6</sup>Typically, PR of  $\sim 50\%$  corresponds to random ranking.



		Pos. MPR (%)		Neg. MPR (%)	
		events	entities	events	entities
Local	Longformer	61.9	59.7	54.8	50.5
	Local CDLM	<b>62.2</b>	<b>60.8</b>	54.6	52.6
Global	Rand CDLM	70.6	69.1	56.6	53.2
	Prefix CDLM	70.7	69.4	58.5	56.5
	CDLM	<b>72.1</b>	<b>70.3</b>	58.0	55.7

Table 6: Cross-document coreference resolution alignment MPR scores of the target span, with respect to the tokens in the same document.

spans. For the positive examples, the gap of up to 10% of MPR between the “Local” and “Global” models shows the advantage of adopting global attention during the pretraining phase. This indicates that the global attention mechanism implicitly helps to encode alignment information.

## 4 Related Work

Recently, long-context language models (Beltagy et al., 2020; Zaheer et al., 2020) introduced the idea of processing multi-document tasks using a single long-context sequence encoder. However, pretraining objectives in these models consider only single documents. Here, we showed that additional gains can be obtained by MLM pretraining using multiple *related* documents as well as a new dynamic global attention pattern.

Processing and aggregating information from multiple documents has been also explored in the context of document retrieval, aiming to extract information from a large set of documents (Guu et al., 2020; Lewis et al., 2020a,b; Karpukhin et al., 2020). These works focus on retrieving relevant information from often a large collection of documents, by utilizing short-context LMs, and then generate information of interest. CDLM instead provides an approach for improving the encoding and contextualizing information across multiple documents. As opposed to the mentioned works, our model utilizes long-context LM and can include broader contexts of more than a single document.

The use of cross-document attention has been recently explored by the Cross-Document Attention (CDA) (Zhou et al., 2020). CDA specifically encodes two documents, using hierarchical attention networks, with the addition of cross attention between documents, and makes similarity decision between them. Similarly, the recent DCS model (Ginzburg et al., 2021) suggested a cross-document finetuning scheme for unsuper-

vised document-pair matching method (processing only two documents at once). Our CDLM, by contrast, is a general pretrained language model that can be applied to a variety of multi-document downstream tasks, without restrictions on the number of input documents, as long as they fit the input length of the Longformer.

Finally, our pretraining scheme is conceptually related to cross-encoder models that leverage simultaneously multiple related information sources. For example, the Translation Language Model (TLM) (Conneau and Lample, 2019) encodes together sentences and their translation, while certain cross-modality encoders pretrain over images and texts in tandem (e.g., ViLBERT (Lu et al., 2019)).

## 5 Conclusion

We presented a novel pretraining strategy and technique for cross-document language modeling, providing better encoding for cross-document (CD) downstream tasks. Our contributions include the idea of leveraging clusters of related documents for pretraining, via cross-document masking, along with a new long-range attention pattern, together driving the model to learn to encode CD relationships. This was achieved by extending the global attention mechanism of the Longformer model to apply already in pretraining, creating encodings that attend to long-range information across and within documents. Our experiments assess that our cross-document language model yields new state-of-the-art results over several CD benchmarks, while, in fact, employing substantially smaller models. Our analysis showed that CDLM implicitly learns to recover long-distance CD relations via the attention mechanism. We propose future research to extend this framework to train larger models, and to develop cross-document sequence-to-sequence models, which would support CD tasks that involve a generation phase.

## Acknowledgments

We thank Doug Downey and Luke Zettlemoyer for fruitful discussions and helpful feedback, and Yoav Goldberg for helping us connect with collaborators on this project. The work described herein was supported in part by grants from Intel Labs, the Israel Science Foundation grant 1951/17, the Israeli Ministry of Science and Technology, and the NSF Grant OIA-2033558.

## References

- Emily Allaway, Shuai Wang, and Miguel Ballesteros. 2021. Sequential cross-document coreference resolution. *arXiv preprint arXiv:2104.08413*.
- Akari Asai, Kazuma Hashimoto, Hannaneh Hajishirzi, Richard Socher, and Caiming Xiong. 2020. Learning to retrieve reasoning paths over wikipedia graph for question answering. In *International Conference on Learning Representations (ICLR)*.
- 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*, volume 1, pages 563–566. Citeseer.
- Shany Barhom, Vered Shwartz, Alon Eirew, Michael Bugert, Nils Reimers, and Ido Dagan. 2019. [Re-visiting 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.
- Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.
- Chandra Bhagavatula, Sergey Feldman, Russell Power, and Waleed Ammar. 2018. [Content-based citation recommendation](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 238–251, New Orleans, Louisiana. Association for Computational Linguistics.
- Arie Cattan, Alon Eirew, Gabriel Stanovsky, Mandar Joshi, and Ido Dagan. 2020. Streamlining cross-document coreference resolution: Evaluation and modeling. *ArXiv*, abs/2009.11032.
- Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. [What does BERT look at? an analysis of BERT’s attention](#). In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.
- Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems (NIPS)*.
- 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. “Bag of Events” approach to event coreference resolution. supervised classification of event templates. *IJCLA*, page 11.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. [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.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. [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.
- Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. [The hitchhiker’s guide to testing statistical significance in natural language processing](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.
- Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. [Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.
- Dvir Ginzburg, Itzik Malkiel, Oren Barkan, Avi Caciularu, and Noam Koenigstein. 2021. [Self-supervised document similarity ranking via contextualized language models and hierarchical inference](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3088–3098, Online. Association for Computational Linguistics.
- Kelvin Guu, Kenton Lee, Zora Tung, Panupong Papat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *Proceedings of the International Conference on Machine Learning (ICML)*.
- Eran Hirsch, Alon Eirew, Ori Shapira, Avi Caciularu, Arie Cattan, Ori Ernst, Ramakanth Pasunuru, Hadar Ronen, Mohit Bansal, and Ido Dagan. 2021. ifacetsum: Coreference-based interactive faceted summarization for multi-document exploration. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations*.
- Jyun-Yu Jiang, Mingyang Zhang, Cheng Li, Michael Bendersky, Nadav Golbandi, and Marc Najork. 2019. Semantic text matching for long-form documents. In *The World Wide Web Conference (WWW)*.

- Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. [Dense passage retrieval for open-domain question answering](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. 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.
- 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.
- Mike Lewis, Marjan Ghazvininejad, Gargi Ghosh, Armen Aghajanyan, Sida Wang, and Luke Zettlemoyer. 2020a. Pre-training via paraphrasing. *Advances in Neural Information Processing Systems (NeurIPS)*.
- Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020b. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- 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*.
- Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. [S2ORC: The semantic scholar open research corpus](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.
- Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems (NIPS)*.
- Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 25–32. Association for Computational Linguistics.
- Yehudit Meged, Avi Caciularu, Vered Shwartz, and Ido Dagan. 2020. [Paraphrasing vs coreferring: Two sides of the same coin](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4897–4907, Online. Association for Computational Linguistics.
- Nafise Sadat Moosavi and Michael Strube. 2016. [Which coreference evaluation metric do you trust? a proposal for a link-based entity aware metric](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 632–642, Berlin, Germany. Association for Computational Linguistics.
- Martin Potthast, Matthias Hagen, Tim Gollub, Martin Tippmann, Johannes Kiesel, Paolo Rosso, Efstathios Stamatatos, and Benno Stein. 2013. Overview of the 5th international competition on plagiarism detection. In *Conference on Multilingual and Multimodal Information Access Evaluation (CLEF)*.
- Dragomir R Radev, Pradeep Muthukrishnan, Vahed Qazvinian, and Amjad Abu-Jbara. 2013. The ACL anthology network corpus. *Language Resources and Evaluation*, 47(4):919–944.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.
- Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020a. [A primer in BERTology: What we know about how BERT works](#). *Transactions of the Association for Computational Linguistics*, 8:842–866.
- Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020b. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics (TACL)*.
- Vered Shwartz, Gabriel Stanovsky, and Ido Dagan. 2017. [Acquiring predicate paraphrases from news tweets](#). In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (\*SEM 2017)*, pages 155–160, Vancouver, Canada. Association for Computational Linguistics.
- Qingying Sun, Zhongqing Wang, Qiaoming Zhu, and Guodong Zhou. 2018. [Stance detection with hierarchical attention network](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2399–2409, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. 2021. Long range arena: A benchmark for efficient transformers. In *International Conference on Learning Representations (ICLR)*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all

- you need. In *Advances in neural information processing systems (NIPS)*.
- Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of the 6th conference on Message understanding*, pages 45–52. Association for Computational Linguistics.
- Sarah Wiegrefe and Yuval Pinter. 2019. [Attention is not not explanation](#). 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 11–20, Hong Kong, China. Association for Computational Linguistics.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Liu Yang, Mingyang Zhang, Cheng Li, Michael Bendersky, and Marc Najork. 2020. Beyond 512 tokens: Siamese multi-depth transformer-based hierarchical encoder for long-form document matching. In *Proceedings of the ACM International Conference on Information & Knowledge Management (CIKM)*.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. [HotpotQA: A dataset for diverse, explainable multi-hop question answering](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.
- Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. [Hierarchical attention networks for document classification](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California. Association for Computational Linguistics.
- Xiaodong Yu, Wenpeng Yin, and Dan Roth. 2020. Paired representation learning for event and entity coreference. *arXiv preprint arXiv:2010.12808*.
- Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big Bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems (NeurIPS)*.
- Yutao Zeng, Xiaolong Jin, Saiping Guan, Jiafeng Guo, and Xueqi Cheng. 2020. [Event coreference resolution with their paraphrases and argument-aware embeddings](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3084–3094, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Chen Zhao, Chenyan Xiong, Corby Rosset, Xia Song, Paul Bennett, and Saurabh Tiwary. 2020. Transformer-XH: Multi-evidence reasoning with extra hop attention. In *International Conference on Learning Representations (ICLR)*.
- Xuhui Zhou, Nikolaos Pappas, and Noah A. Smith. 2020. [Multilevel text alignment with cross-document attention](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5012–5025, Online. Association for Computational Linguistics.



## A Dataset Statistics and Details

In this section, we provide details regarding the pretraining corpus and benchmarks we used during our experiments.

### A.1 Multi-News Corpus

We used the preprocessed, not truncated version of Multi-News, which totals 322MB of uncompressed text.<sup>7</sup> Each one of the preprocessed documents contains up to 500 tokens. The average and 90<sup>th</sup> percentile of input length is 2.5k and 3.8K tokens, respectively. In Table 7 we list the number of related documents per cluster. This follows the original dataset construction suggested in Fabbri et al. (2019).

# of docs in cluster	Frequency
3	12,707
4	5,022
5	1,873
6	763
7	382
8	209
9	89
10	33
<b>Total</b>	21,078

Table 7: MultiNews training set statistics.

### A.2 ECB+ Dataset

In Table 8, we list the statistics about training, development, and test splits regarding the topics, documents, mentions and coreference clusters. We follow the data split used by previous works (Cybulska and Vossen, 2015; Kenyon-Dean et al., 2018; Barhom et al., 2019): For training, we consider the topics: 1, 3, 4, 6-11, 13-17, 19-20, 22, 24-33; For Validation, we consider the topics: 2, 5, 12, 18, 21, 23, 34, 35; For test, we consider the topics: 36-45.

	Train	Validation	Test
Topics	25	8	10
Docs	594	196	206
Mentions	3808/4758	1245/1476	1780/2055
Clusters	411/472	129/125	182/196

Table 8: ECB+ dataset statistics. The slash numbers for Mentions and Clusters represent event/entity statistics.

<sup>7</sup>We used the dataset available in <https://drive.google.com/open?id=1qZ3zJBv0zrUy4HVWxnx33IsrHGimXLPy>.

### A.3 Paper Citation Recommendation & Plagiarism Detection Datasets

In Table 9, we list the statistics about training, development, and test splits for each benchmark separately, and in Table 10, we list the document and example counts for each benchmark. The statistics are taken from Zhou et al. (2020).

Dataset	Train	Validation	Test
AAN	106,592	13,324	13,324
OC	240,000	30,000	30,000
S2ORC	152,000	19000	19000
PAN	17,968	2,908	2,906

Table 9: Document-to-Document benchmarks statistics: Details regarding the training, validation, and test splits.

Dataset	# of doc pairs	# of docs
AAN	132K	13K
OC	300K	567K
S2ORC	190K	270K
PAN	34K	23K

Table 10: Document-to-Document benchmarks statistics: The reported numbers are the count of document pairs and the count of unique documents.

The preprocessing of the datasets performed by Zhou et al. (2020) includes the following steps: For AAN, only pairs of documents that include abstracts are considered, and only their abstracts are used. For OC, only one citation per paper is considered, and the dataset was downsampled significantly. For S2ORC, formed pairs of citing sections and the corresponding abstract in the cited paper are included, and the dataset was downsampled significantly. For PAN, pairs of relevant segments out of the entire document were extracted.

For all the datasets, negative pairs were sampled randomly. Then, a standard preprocessing that includes filtering out characters that are not digits, letters, punctuation, or white space in the texts was performed.

## B CDLM Pretraining Hyperparameters

In this section, we detail the hyperparameters setting of the models we pretrained, including CDLM Prefix CDLM, Rand CDLM, and Local CDLM: The input sequences are of the length of 4,096, effective batch size of 64 (using gradient accumulation and batch size of 8), a maximum learning rate

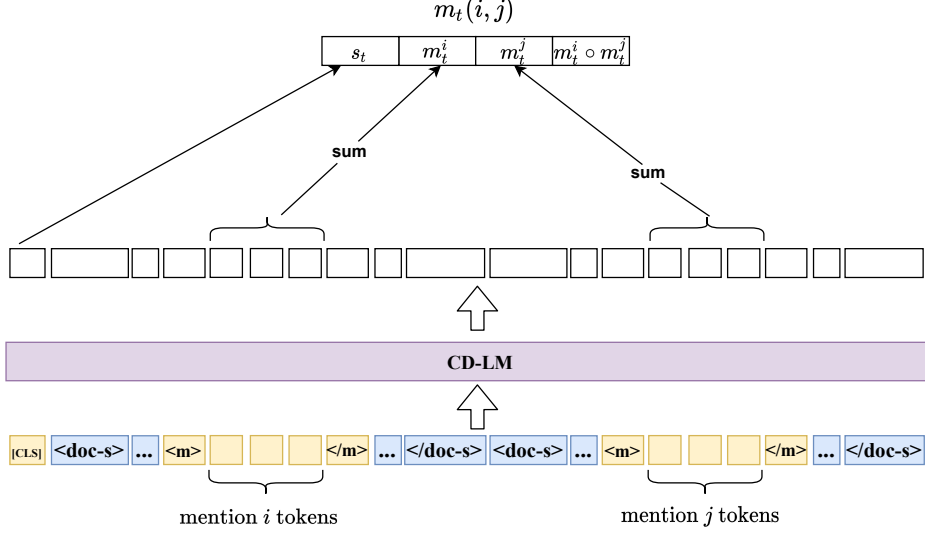


Figure 4: CD-coreference resolution pairwise mention representation, using the new setup, for our CDLM models.  $m_t^i, m_t^j$  and  $s_t$  are the cross-document contextualized representation vectors for mentions  $i$  and  $j$ , and of the [CLS] token, respectively.  $m_t^i \circ m_t^j$  is the element-wise product between  $m_t^i$  and  $m_t^j$ .  $m_t(i, j)$  is the final produced pairwise-mention representation. The tokens colored in yellow represent global attention, and tokens colored in blue represent local attention.

of  $3e-5$ , and a linear warmup of 500 steps, followed by a power 3 polynomial decay. For speeding up the training and reducing memory consumption, we used the mixed-precision (16-bits) training mode. The pretraining took 8 days, using eight 48GB RTX8000 GPUs. The rest of the hyperparameters are the same as for RoBERTa (Liu et al., 2019). Note that training CDLM using the large version of the Longformer model might require 2-3 times more memory and time.

## C Finetuning on Downstream Tasks

In this section, we elaborate further implementation details regarding the downstream tasks that we experimented, including the hyperparameter choices and the algorithms used.

### C.1 Cross-Document Coreference Resolution

The setup for our cross-document coreference resolution pairwise scoring is illustrated in Figure 4. We concatenate the relevant documents using the special document separator tokens, then encode them using our CDLM along with the [CLS] token at the beginning of this sequence, as suggested in Section 2.2. For within-document coreference candidate examples, we use just the single containing document with one set of document separators, for the single input document. Inspired by Yu et al. (2020), we use candidate mention marking:

we wrap the mentions with special tokens  $\langle m \rangle$  and  $\langle /m \rangle$  in order to direct the model to specifically pay attention to the candidates representations. Additionally, we assign global-attention to [CLS],  $\langle m \rangle$ ,  $\langle /m \rangle$ , and the mention tokens themselves, according to the finetuning strategy proposed in Section 2.2. Our final pairwise-mention representation is formed like in Zeng et al. (2020) and Yu et al. (2020): We concatenate the cross-document contextualized representation vectors for the  $t^{\text{th}}$  sample:

$$m_t(i, j) = [s_t, m_t^i, m_t^j, m_t^i \circ m_t^j],$$

where  $[\cdot]$  denotes the concatenation operator,  $s_t$  is the cross-document contextualized representation vector of the [CLS] token, and each of  $m_t^i$  and  $m_t^j$  is the sum of candidate tokens of the corresponding mentions ( $i$  and  $j$ ). Then, we train the pairwise scorer according to the suggested finetuning scheme. At test time, similar to most recent works, we apply agglomerative clustering to merge the most similar cluster pairs.

Regarding the training data collection and hyperparameter setting, we adopt the same protocol as suggested in Cattani et al. (2020).<sup>8</sup> Our training set is composed of positive instances which consist of all the pairs of mentions that belong to the same

<sup>8</sup>We used the implementation taken from [https://github.com/ariecattan/cross\\_encoder](https://github.com/ariecattan/cross_encoder)

coreference cluster, while the negative examples are randomly sampled.

The resulting feature vector is passed through a MLP pairwise scorer that is composed of one hidden layer of the size of 1024, followed by the Tanh activation. We finetune our models for 10 epochs, with an effective batch size of 128. We used eight 32GB V100-SMX2 GPUs for finetuning our models. The finetuning process took  $\sim 28$  and  $\sim 45$  hours per epoch, for event coreference and entity coreference, respectively.

## C.2 Multi-Document Classification

We tune our models for 8 epochs, using a batch size of 32, and used the same hyperparameter setting from Zhou et al. (2020, Section 5.2).<sup>9</sup> We used eight 32GB V100-SMX2 GPUs for finetuning our models. The finetuning process took  $\sim 2, \sim 5, \sim 3$ , and  $\sim 0.5$  hours per epoch, for AAN, OC, S2ORC, and for PAN, respectively. We used the mixed-precision training mode, to reduce time and memory consumption.

## C.3 Multihop Question Answering

For preparing the data for training and evaluation, we follow our finetuning scheme: for each example, we concatenate the question and all the 10 paragraphs in one long context. We particularly use the following input format with special tokens and our document separators: “[CLS] [q] question [/q] <doc-s><t> title<sub>1</sub> </t> <s> sent<sub>1,1</sub> </s> <s> sent<sub>1,2</sub> </s> </doc-s> ... <t> <doc-s> title<sub>2</sub> </t> sent<sub>2,1</sub> </s> <s> sent<sub>2,2</sub> </s> <s> ...” where [q], [/q], <t>, </t>, <s>, </s>, [p] are special tokens representing, question start and end, paragraph title start and end, and sentence start and end, respectively. The new special tokens were added to the models vocabulary and randomly initialized before task finetuning. We use global attention to question tokens, paragraph title start tokens as well as sentence tokens. The model’s structure is taken from Beltagy et al. (2020).

Similar to Beltagy et al. (2020), we finetune our models for 5 epochs, using a batch size of 32, learning rate of  $1e-4$ , 100 warmup steps. Finetuning on our models took  $\sim 6$  hours per epoch, using four 48GB RTX8000 GPUs for finetuning our models.

---

<sup>9</sup>we used the script [https://github.com/XuhuiZhou/CDA/blob/master/BERT-HAN/run\\_ex\\_sent.sh](https://github.com/XuhuiZhou/CDA/blob/master/BERT-HAN/run_ex_sent.sh)