Cross-document Event Coreference Search: Task, Dataset and Modeling

Alon Eirew[1,2] **Avi Caciularu**[1] **Ido Dagan**[1]
[1]Bar Ilan University, Ramat-Gan, Israel [2]Intel Labs, Israel
alon.eirew@intel.com
avi.c33@gmail.com
dagan@cs.biu.ac.il

Abstract

The task of Cross-document Coreference Resolution has been traditionally formulated as requiring to identify *all* coreference links across a given set of documents. We propose an appealing, and often more applicable, complementary set up for the task – *Cross-document Coreference Search*, focusing in this paper on event coreference. Concretely, given a mention in context of an event of interest, considered as a query, the task is to find all coreferring mentions for the query event in a large document collection. To support research on this task, we create a corresponding dataset, which is derived from Wikipedia while leveraging annotations in the available Wikipedia Event Coreference dataset (WEC-Eng). Observing that the coreference search setup is largely analogous to the setting of Open Domain Question Answering, we adapt the prominent Deep Passage Retrieval (DPR) model to our setting, as an appealing baseline. Finally, we present a novel model that integrates a powerful coreference scoring scheme into the DPR architecture, yielding improved performance.

1 Introduction

Cross-Document Event Coreference (CDEC) resolution is the task of identifying clusters of text mentions, across multiple texts, that refer to the same event. For example, consider the following two underlined event mentions from the WEC-Eng CDEC dataset (Eirew et al., 2021):

1. *...On **14 April 2010**, an earthquake struck the prefecture, registering a magnitude of 6.9 (USGS, EMSC) or 7.1 (Xinhua). It originated in the **Yushu Tibetan** Autonomous Prefecture...*

2. *...a school mostly for **Tibetan** orphans in Chindu County, Qinghai, after the **2010 Yushu** earthquake destroyed the old school...*

Both event mentions refer to the same earthquake, as can be determined by the shared event arguments (2010, Yushu, Tibetan). In event coreference resolution, the goal is to cluster event mentions that refer to the same event, whether within a single document or across a document collection.

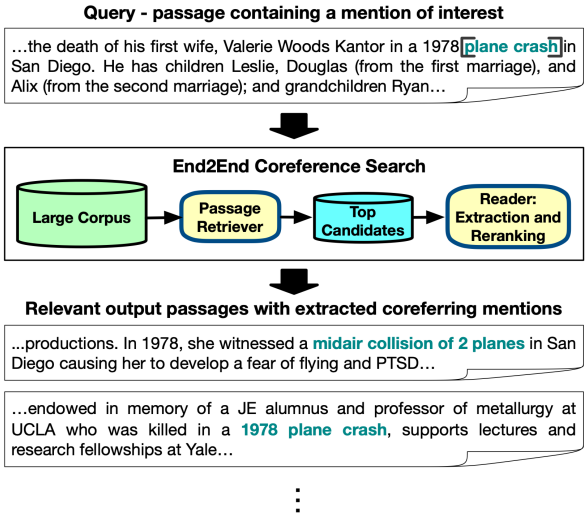

Figure 1: Example of Coreference Search. Provided with a query passage containing a mention of interest, a coreference search system retrieves from a large corpus the best candidate passages containing mentions coreferring with the query.

Currently, with the growing number of documents describing real-world events and event-oriented information, the need for efficient methods for accessing such information is apparent. Successful and efficient identification, clustering, and access to event-related information, may be beneficial for a broad range of applications at the multi-text level, that need to match and integrate information across documents, such as multi-document summarization (Falke et al., 2017; Liao et al., 2018), multi-hop question answering (Dhingra et al., 2018; Wang et al., 2019) and Knowledge Base Population (KBP) (Lin et al., 2020).

Currently, the CDEC task, as formed in corresponding datasets, is intended at creating models that exhaustively resolve all coreference links in

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 900 - 913
December 7-11, 2022

a given dataset. However, an applicable realistic scenario may require to efficiently search and extract coreferring events of only specific events of interest. A typical such use-case can be of a user reading a text and encountering an event of interest (for example, the *plane crash* event in Figure 1), and then wishing to further explore and learn about the event from a large document collection.

To address such needs, we propose an appealing, and often more applicable, complementary set up for the task – *Cross-document Coreference Search* (Figure 1), focusing in this paper on event coreference. Concretely, given a mention in context of an event of interest, considered as a query, the task is to find all coreferring mentions for the query event in a large corpus.

Such coreference resolution search use-case cannot be addressed currently, for two main reasons: (1) Existing CDEC datasets are relatively small for the realistic representation of a search task; (2) Current CDEC models, which are designed at linking all coreference links in a given dataset, are inapplicable in terms of computation at the much larger search space required by realistic coreference resolution search scenarios.

To facilitate research on this setup, we present a large dataset, derived from Wikipedia, by leveraging existing annotations in the Wikipedia Event Coreference dataset (WEC) (Eirew et al., 2021). Our curated dataset resembles in structure to an Open-domain QA (ODQA) dataset (Berant et al., 2013; Baudiš and Šedivý, 2015; Joshi et al., 2017; Kwiatkowski et al., 2019; Rajpurkar et al., 2016), containing a set of coreference queries and a large passage collection for retrieval.

Observing that the coreference search setup is largely analogous to the setting of Open Domain Question Answering, we adapt the prominent Deep Passage Retrieval (DPR) model to our setting, as an appealing baseline. Further, motivated to integrate coreference modeling into DPR, we adapted components inspired by a prominent within-document end-to-end coreference resolution model (Lee et al., 2017), which was previously applied also to the CDEC task (Cattan et al., 2020). Thus, we developed an integrated model that leverages components from both DPR and the coreference model of Lee et al. (2017). Our novel model yields substantially improved performance on several important evaluation metrics.

Our dataset[1] and code[2] are released for open access.

2 Background

In this section, we first describe the Cross Document Event Coreference (CDEC) task, datasets and models (§2.1) and then review the common open-domain QA model architecture (§2.2).

2.1 Cross-Document Event Coreference Resolution

ECB+ (Cybulska and Vossen, 2014) is the most commonly used dataset for training and testing models for cross-document event coreference resolution. This corpus consists of documents partitioned into 43 clusters, each corresponding to a certain news topic. ECB+ is relatively small, where on average only 1.9 sentences per document were selected for annotation, yielding only 722 non-singleton coreference clusters in total (that is, clusters containing more than a single event mention, while singleton clusters correspond to mentions that do not hold a coreference relation with any other mention in the data).

Since annotating a CDEC dataset is a very challenging task, several annotation methods try to semi-automatically create a CDEC dataset by taking advantage of available resources. The Gun Violence Corpus (GVC) (Vossen et al., 2018) leveraged a structured database recording gun violence events for creating an annotation scheme for gun violence related events. In total GVC annotated 7,298 mentions distributed into 1,046 non-singleton clusters.

More recently, WEC-Eng (Eirew et al., 2021) and HyperCoref (Bugert and Gurevych, 2021) leveraged article hyperlinks pointing to the same concept in order to create an automatic annotation process. This annotation scheme helped HyperCoref curate 2.7M event mentions distributed among 0.8M event clusters, extracted from news articles. The smaller WEC-Eng curates 43,672 event mentions distributed among 7,597 non-singleton clusters. Differently then HyperCoref, the WEC-Eng development set (containing 1,250 mentions and 233 clusters) and test set (contains 1,893 mentions and 322 clusters) have gone through a manual validation process (see Table 1), ensuring their high quality.

[1] https://huggingface.co/datasets/Intel/CoreSearch

[2] https://github.com/AlonEirew/CoreSearch

All the above mentioned datasets are targeted for models which exhaustively resolve all coreference links within a given dataset (Barhom et al., 2019; Meged et al., 2020; Cattan et al., 2020; Caciularu et al., 2021; Yu et al., 2020; Held et al., 2021; Allaway et al., 2021; Hsu and Horwood, 2022). This setting resembles the within-document coreference resolution setting, where similarly all links are exhaustively resolved in a given single-document. However, while within-document coreference resolution is contained to a single document, CDCR might relate to an unbounded multi-text search space (e.g., news articles, Wikipedia articles, court and police records and so on). To that end, we aim at a task and dataset for modeling CDEC as a search problem. To facilitate a large corpus for a realistic representation of such a task, while ensuring reliable development and test sets, we adopted the WEC-Eng[3] as the basis for our dataset creation (§3).

Within Document Coreference Resolution Recent within-document coreference resolution models (Lee et al., 2018; Joshi et al., 2019; Kantor and Globerson, 2019; Wu et al., 2020), were inspired by the end-to-end model architecture introduced by Lee et al. (2017). In particular, two distinct components were adopted in those works, which were shown to be effective in detecting mentions and their coreference relations, both in the within-document and cross-document (Cattan et al., 2020) settings. In our proposed model, we similarly adopt those two components to better represent coreference relations, in the coreference search settings.

2.2 Open-Domain Question Answering

Open-domain question answering (ODQA) (Voorhees, 1999), is concerned with answering factoid questions based on a large collection of documents. Modern open-domain QA systems have been restructured and simplified by combining information retrieval (IR) techniques and neural reading comprehension models (Chen et al., 2017). In those approaches, a retriever component finds documents that might contain an answer from a large collection of documents, followed by a reader component that finds a candidate answer in a given document (Lee et al., 2019; Yang et al., 2019; Karpukhin et al., 2020).

[3]The larger magnitude of HyperCoref makes it a suitable candidate for our CoreSearch. However, since HyperCoref is not publicly released, we could not evaluate on it. We leave this part to future work.

	Mentions	None-Singleton Clusters
WEC-Eng (train)	40,529	7,042
WEC-Eng (dev)	1,250	216
WEC-Eng (test)	1,893	306

Table 1: WEC-Eng Dataset Statistics. **Mentions**: The total number of event mentions within the corresponding section. **Non-Singleton Clusters**: Number of event clusters containing more than a single event mention.

	Train	Dev	Test	Total
WEC-Eng Validated Data				
# Clusters	237	49	236	522
# Passages (with Mentions)	1,503	341	1,266	3,110
# Added Destructor Passages	922,736	923,376	923,746	2,769,858
# Total Passages	924,239	923,717	925,012	2,772,968

Table 2: CoreSearch dataset statistics.

We observe that the Cross-Document Event Coreference Search (CDES) setting resembles the ODQA task. Specifically, given a passage containing a mention of interest, considered as a *query*, CDES is concerned with finding mentions coreferring with the query event in a large document collection. To facilitate research in this task, we created a dataset similar in structure to ODQA datasets (Berant et al., 2013; Baudiš and Šedivý, 2015; Joshi et al., 2017; Kwiatkowski et al., 2019; Rajpurkar et al., 2016), and established a suitable model resembling in architecture to the recent two-step (retriever/reader) systems, as described in the following sections.

3 The CoreSearch Dataset

We formulated the *Cross-Document Event Coreference Search* task following a similar approach to open-domain question answering (illustrated in Figure 1). Specifically, given a query containing a marked target event mention, along with a passage collection, the goal is to retrieve all the passages from the passage collection that contain an event mention coreferring with the query event, and extract the coreferring mention span of each retrieved passage.

To facilitate research on this task, we present a large dataset, derived from Wikipedia, termed *CoreSearch*. In this section we describe the CoreSearch dataset structure (§3.1), following by describing the structure of a single query instance (§3.2).

3.1 Dataset Structure

The CoreSearch dataset consists of two separate passage collections: (1) a collection of passages containing manually annotated coreferring event mention, and (2) a collection of destructor passages.

Annotated Data The CoreSearch passage collection which contains manually annotated event mentions was created by importing the validated portion of the WEC-Eng (Eirew et al., 2021) dataset (§2.1 and Table 1).

Specifically, we merged the WEC-Eng validated test and development set coreference clusters into a single collection of 522 none-singleton clusters ("Non-Singleton Culsters" in Table 1 and "Clusters" in Table 2). We then split the clusters between CoreSearch train, development and test sets. Each cluster contains passages that form our annotated passage collection.

Those passages will serve the roles of queries and of positive retrieved coreferring passages.

Destructor Passages In order to collect a large collection of passages for challenging and realistic retrieval, we generate negative passages (i.e., destructing passages) using two resources: (1) The entire WEC-Eng train set, which is not manually validated, though quite reliable; (2) By extracting the first paragraph of any Wikipedia article not containing a hyperlink to any of the CoreSearch annotated passages, and hence are unlikely to corefer with any of them (Table 2).

Cluster Types We observe that our annotated data is characterized by two prominent types of coreference clusters: *Type-1* - clusters containing only passages with event mention spans that include the event time or location (e.g., "2006 Dahab bombings", "2013's BET Awards"), and *Type-2* - clusters that are comprised partly of passages as in Type-1, as well as passages containing mention spans without any event identifying participants (e.g., "the deadliest earthquake on record", "BET Awards", "plane crash"). Naturally, Type-2 clusters will create queries/passage examples with a higher degree of difficulty. Identifying coreference for Type-2 clusters is indeed challenging in our dataset, because WEC-Eng includes a multitude of event mentions which are similar lexically but do not corefer (e.g., different earthquakes) (Eirew et al., 2021), requiring a model to identify event coreference using the arguments in the surrounding context.

Query	Unique Positive Passage Mention Answers
...On 14 April 2010, an **earthquake** struck the prefecture, registering a magnitude of 6.9...	'Yushu earthquake', '2010 Yushu earthquake', 'earthquake in Qinghai', 'earthquake in 2010', 'Qinghai earthquake'
...2012–13 season 10–21, 6–12 in MAAC play to finish in eighth place. They lost in the first round of the **MAAC Tournament** to...	'MAAC Tournament', '2013 MAAC Tournament'
...Salo and the band The Ark won Melodifestivalen 2007 and went on to represent Sweden in the **Eurovision Song Contest** 2007 with the song...	'52nd Eurovision Song Contest 2007', 'previous contest', '2007 edition of the Contest', '2007 contest', 'that year's contest', 'Eurovision 2007'
...finished the season 18–15, 10–8 in Pac-10 play. They lost to USC in the quarterfinals **Pac-10 tournament**...	'2011 Pacific-10 Conference Men's Basketball Tournament', '2011 Pac-10 Tournament', '2011 Pac-10 tournament'
...The film was planned to premiere at the 65th annual **Cannes International Film Festival** in May 2012, but in late 2011	'Cannes Short Film Corner', '2012 Cannes Film Festival', 'Cannes Film Festival', '65th Annual Cannes Film Festival'

Table 3: Sample of five queries containing a mention (highlighted in **green**) without event participants, and the corresponding cluster mentions (blue highlights passages mentions without event participants), illustrating challenging query examples in CoreSearch dataset.

To measure the distribution of cluster types within CoreSearch, we randomly sampled 20 clusters and found 90% are of type-2, demonstrating the challenging nature of the CoreSearch data. Table 3 illustrates examples of queries extracted randomly from five type-2 clusters.

3.2 CoreSearch Instance Structure

An instance in the CoreSearch dataset is comprised of: (1) a query passages pulled from the annotated passage collection; (2) The collection of all other passages, which are considered as the passage collection for retrieval. Passages in the passage collection which belong to the same cluster as the pulled query are considered positive passages, while all the rest as negative passages.

Potential Language Adaptation The CoreSearch dataset is built on top of the English version of WEC (WEC-Eng). Consequently, since WEC is adoptable to other languages with relatively low effort (Eirew et al., 2021), and the process for deriving CoreSearch from it is simple and fully automatic, the CoreSearch dataset may be adopted to other languages as well with very similar effort (as

for WEC).

4 Coreference-search Models

In this section, we aim to devise an effective baseline for our event coreference search task to be trained on our dataset. Following the observation that coreference search formulation resembles the open-domain QA (ODQA) (§2.2), we propose an end-to-end neural architecture, comprised of a *retriever* and a *reader* models. Given a query passage, the retriever selects the top-k most relevant passage candidates out of the entire passage corpus (§4.1). Then, the reader is responsible for re-ranking the retrieved passages and extracting the coreferring event span, by using a reading comprehension module (§4.2).

4.1 The Retriever Model

Given a query passage containing an event mention of choice, the goal of the *retriever* is to select the top-k relevant passage candidates out of a large collection of passages. To that end, we build upon the foundations of the Dense Passage Retriever model (Karpukhin et al., 2020) and employ a similar retriever.

Similarly to DPR, we propose to encode the query passage $q_i = [\texttt{CLS}, q_i^1, \ldots, q_i^{n_i}]$ and a candidate passage $p_j = [\texttt{CLS}, p_j^1, \ldots, p_j^{n_j}]$ using two distinct neural encoders, $E_Q(\cdot)$ and $E_P(\cdot)$,[4] for mapping their tokens into d-dimensional dense vectors, $[\mathbf{q}_i^{\texttt{CLS}}, \mathbf{q}_i^1, \ldots, \mathbf{q}_i^{n_i}]$ and $[\mathbf{p}_j^{\texttt{CLS}}, \mathbf{p}_j^1, \ldots, \mathbf{p}_j^{n_j}]$ for q_i and p_j, respectively. Here, both $\mathbf{q}_i^{\texttt{CLS}}$ and $\mathbf{p}_j^{\texttt{CLS}}$ denote the last hidden layer contextualized `[CLS]` token representations of q_i and p_j respectively, which are then fed to a dot-product similarity scoring function, which determines candidate passage ranking:

$$\text{sim}(q_i, p_j) = \mathbf{q}_i^{\texttt{CLS}} \cdot \mathbf{p}_j^{\texttt{CLS}} \quad (1)$$

Event Mention Marking In order to accommodate our setup of mention-directed search and to better signal the model to be aware of the query event mention, we edit the query by marking the span of the mention within the query passage by using boundary tokens. Given the query event mention span $m_i = [q_i^k, q_i^{k+1}, \ldots, q_i^{k+l-1}]$, we append the boundary tokens to obtain the final edited query (m_i denotes the sequence of the mention's tokens):

$$q_i = [\texttt{CLS}, \ldots, q_i^{k-1}, \langle \texttt{S} \rangle, m_i, \langle \backslash \texttt{S} \rangle, q_i^{k+l}, \ldots q_i^{n_i}].$$

[4] As in DPR, after training these encoders, we use $E_P(\cdot)$ to build an index for all the passages in the corpus prior to applying the test-time retrieval.

Improved Span Representation For implementing the text encoders $E_Q(\cdot)$ and $E_P(\cdot)$, we employed the SpanBERT[5] (Joshi et al., 2020) model as our query and passage encoders. SpanBERT is an appealing encoder, as it was pre-trained for better span representations, rather then the individual tokens, and was also shown to be more effective for coreference resolution tasks (Joshi et al., 2020; Wu et al., 2020).

During our preliminary experiments, we observed that both the additional event mention marking as well as replacing BERT with SpanBERT contributed significantly to the performance over our dataset.

Positive and Negative Training Examples We construct our positive and negative examples by iterating sequentially through every training set event coreference cluster $C_j = [m_1, m_2, \ldots, m_{|C_j|}]$, where m_i denotes an event mention surrounded with its context (the entire passage). Given each event mention m_i acting as a query q_i, we construct one positive coreference example for each of the remaining $|C_j| - 1$ coreferring event mentions in the cluster. Then, for each such positive example, we first construct one "challenging" negative example by selecting randomly one of the top-20 passages returned by the BM25 retrieval model for the corresponding query. In addition, for each query in a training batch, we create additional ("easier") in-batch negative examples by taking the "challenging" passages of all other queries in the current batch, similarly to Karpukhin et al. (2020).

Objective Let $\mathcal{D} = \langle q_i, p_i^+, p_{i,1}^-, \ldots, p_{i,n}^- \rangle_{i=1}^m$ be the CoreSearch training set. Similarly to Karpukhin et al. (2020), the goal is to optimize the negative log likelihood loss of the positive passage, which is based on the contrastive loss:

$$\mathcal{L}\left(q_i, p_i^+, p_{i,1}^-, \ldots, p_{i,n}^-\right) = -\log \frac{e^{\text{sim}(q_i, p_i^+)}}{e^{\text{sim}(q_i, p_i^+)} + \sum_{j=1}^n e^{\text{sim}(q_i, p_i^-)}}. \quad (2)$$

4.2 The Reader Model

Given a mention surrounded by its context as the query, and its top-k retrieved passages, the reader model is tasked to (1) re-rank the retrieved passages according to a passage selection score and (2) extract the candidate mention span from each passage.

[5] DPR originally used BERT (Devlin et al., 2019) as their query and passage encoders.

We implemented two flavours of readers, a DPR baseline (§4.2.1), and a DPR reader enhanced with event coreference scores (§4.2.2).

4.2.1 DPR Reader Baseline

We implemented a DPR-based passage selection model that acts as re-ranker through cross-encoding the query and the passage. Specifically, we append a query q_i (including the event mention marker tokens, see §4.1) and a passage p_j, and feed the concatenated input sequence to the RoBERTa text encoder $E_R(\cdot)$ (Liu et al., 2019). Similarly to Karpukhin et al. (2020), we then use the output (last hidden layer) token representations to predict three probability distributions. We compute the span score of the s^{th} to t^{th} tokens from the j^{th} passage as $P_{\text{start},j}(s) \times P_{\text{end},j}(t)$, and a passage selection score of the j^{th} passage as $P_{\text{select}}(j)$:

$$P_{\text{start},j}(s) = \text{softmax}(\mathbf{P}_j \mathbf{w}_{\text{start}})_s \quad (3)$$

$$P_{\text{end},j}(t) = \text{softmax}(\mathbf{P}_j \mathbf{w}_{\text{end}})_t \quad (4)$$

$$P_{\text{select}}(j) = \text{softmax}(\hat{\mathbf{P}}^T \mathbf{w}_{\text{select}})_j, \quad (5)$$

where $[\cdot]$ denotes column concatenation, $\mathbf{P}_j = [\mathbf{p}_j^{\texttt{CLS}}, \mathbf{p}_j^1, \ldots, \mathbf{p}_j^{n_j}]$, $\hat{\mathbf{P}} = [\mathbf{p}_1^{\texttt{CLS}}, \ldots, \mathbf{p}_k^{\texttt{CLS}}]$, k is the number of the retrieved passages, and $\mathbf{w}_{\text{start}}, \mathbf{w}_{\text{end}}, \mathbf{w}_{\text{select}}$ are learned vectors.

4.2.2 Integrating the Coreference Signal

While the above DPR-based reader yields appealing performance (§5.3), we conjecture that the passage selection (Eq. 5), which is based on the passages' `[CLS]` token representations, is sub-optimal for coreference resolution. These representations learn high-quality sentence- or document-level features (Devlin et al., 2019), however in our setting, more fine-grained features are required in order to capture information for better modeling coreference relations between mention spans. Motivated by this hypothesis, we replaced the passage selection component (Eq. 5) with a method adapted from recent neural within-document coreference models (Lee et al., 2017, 2018; Joshi et al., 2019; Kantor and Globerson, 2019; Wu et al., 2020).

Specifically, we aim to model the probability of passage j to be selected by the likelihood it contains an event mention m_j that corefers to the query's event mention m_i:

$$P_{select}(j) = \frac{e^{s(m_j, m_i)}}{\sum_{j=1}^{k} e^{s(m_j, m_i)}} \quad (6)$$

$$s(m_j, m_i) = s_m(m_j) + s_a(m_j, m_i) \quad (7)$$

$$s_m(m_j) = \boldsymbol{w_m} \cdot \texttt{FFNN}_m(\mathbf{g}_j) \quad (8)$$

$$s_a(m_j, m_i) = \boldsymbol{w_a} \cdot \texttt{FFNN}_a([\mathbf{g}_i, \mathbf{g}_j, \mathbf{g}_i \circ \mathbf{g}_j]) \quad (9)$$

Where $s_m(m_j)$ is the mention scorer, $s_a(m_j, m_j)$ is the antecedent scorer that computes coreference likelihood for the pair of mentions, $\circ$ represents the element-wise product of $\mathbf{g}_j$ and $\mathbf{g}_i$, $\mathbf{g}_x = [m_{x,s}, m_{x,t}]$ is the concatenated vector of the first and last token representations of the mention in the passage $x \in \{i, j\}$, and $s(m_i, m_j)$ is the final pairwise score. `FFNN` represents a feed forward neural network with a single hidden layer. Note that standard coreference resolution methods compute also $s_m(m_i)$, however since in our setup the query mention is constant, it can be omitted.

During training, we extract the gold start/end embeddings of the candidate passage, while at inference time, we use the scores computed by Eq. 3 and Eq. 4 (see §4.2.1) in order to extract the most probable plausible mention spans. Invalid spans, who's end precedes their start position point or are longer than a threshold L, are filtered. For the query event mention, we use the same mention marking strategy used for the query encoder (§4.1). We further show in §5.3 that this marking improves the performance of the reader.

5 Experiments and Results

5.1 Implementation Details

Retriever We train the two separate encoders using a maximum query size of 64 tokens for the query encoder. In order to cope with memory constrains, we limit the maximum passage size given to the passage encoder to 180 tokens. Batch size is set to 64. We train our model using four 12GB Nvidia Titan-Xp GPUs.[6]

Reader We train the single cross-encoder using a maximum sequence size of 256 tokens, in order to cope with memory constraints. We use up to 64 tokens from the surrounding query mention context (which in many cases take less then 64 tokens) for query representation, and concatenate the passage

[6]We leveraged and modified the implementations in the Haystack framework of the DPR and BM25 models: https://github.com/deepset-ai/haystack

context using the remaining available sequence. In case the passage context length exceeds available sequence size for passage representation, we segment the passage using overlapping strides of 128 tokens, creating additional passage instances with the same query. The batch size is set to 24, and both $\texttt{FFNN}_m$, $\texttt{FFNN}_a$ use a single hidden layer set to 128. We train the models using two 12GB Nvidia Titan-Xp GPUs.

Hyperparameters All models parameters are updated by the AdamW (Loshchilov and Hutter, 2019) optimizer, with a learning rate set to 10^{-5} and a wight-decay rate of 0.01. We also apply a linear scheduling with warm-up (for 10% of optimization steps) and dropout rate of 0.1. We train all models for 5 epochs and consider the best performing ones over the development set. At inference, we set the retriever top-k parameter to 500.

5.2 Evaluation Measures

In all our experiments, we followed the common evaluation practices used for evaluating Information Retrieval (IR) models (Khattab and Zaharia, 2020; Xiong et al., 2021; Hofstätter et al., 2021; Thakur et al., 2021). Accordingly, we used the following metrics:

Mean Reciprocal Rank (MRR@k) Following common evaluation practices, we set k to 10, expecting that the topmost correct result should appear amongst the top 10 results (that is, no credit is given if the topmost correct result is ranked lower than 10).

Recall (R@k) We report recall at $k \in \{10, 50\}$ for the end-to-end model evaluation, assessing recall in two prototypical cases where the user might choose to look at rather few or rather many results. For the retriever model we report recall at $k \in \{10, 100, 500\}$, illustrating the motivation for the $k = 500$ cutoff point that we chose (beyond which there were no substantial recall gains).

mean Average Precision (mAP@k) The mAP metric assesses the ranking quality of **all** correct results within the top-k ones, measured for $k \in \{10, 50\}$, as measured for recall.[7]

[7]We use mAP rather than Normalized Discounted Cumulative Gain (NDCG), because the latter requires a scaled gold relevancy score for each query result. mAP applies a similar ranking evaluation criterion, but is suitable for binary relevancy scores, which is the case in our coreference setting.

Model	MRR@10	mAP@10	mAP@50	R@10	R@100	R@500
Development						
BM25	57.32	25.92	31.05	27.63	56.09	74.75
Retriever-B$^-$	23.56	4.91	7.71	9.53	42.66	65.09
Retriever-S$^-$	75.4	37.6	44.09	40.06	71.65	86.21
Retriever-S$^+$	**80.92**	**40.43**	**47.5**	**41.59**	**74.03**	**87.53**
Test						
BM25	62.45	29.75	34.02	27.82	57.9	74.75
Retriever-S$^+$	**69.1**	**35.73**	**43.24**	**37.44**	**75.31**	**87.12**

Table 4: Retriever results on CoreSearch development and test sets. **BM25**: BM25 score; **Retriever-B$^-$**: DPR retriever using BERT, without mention boundary tokens; **Retriever-S$^-$**: DPR retriever using SpanBERT, without boundary tokens; **Retriever-S$^+$**: Our complete retriever, with boundary tokens

Model	MRR@10	mAP@10	mAP@50	R@10	R@50	EM	F1
Development							
E2E-DPR$^-$	89.81	58.19	68.35	55.16	**84.47**	74.99	82.97
E2E-DPR$^+$	92.48	58.23	65.14	53.34	73.81	79.59	**88.91**
E2E-Integrated	**94.05**	**61.82**	**70.81**	**56.53**	82.19	**83.31**	88.78
Test							
E2E-DPR$^-$	87.93	60	70.51	52.3	**86.62**	65.35	77.77
E2E-DPR$^+$	88.18	58.26	66.37	49.24	76.66	69.87	82.92
E2E-Integrated	**90.06**	**63.26**	**72.91**	**53.5**	84.35	**71.44**	**84.16**

Table 5: End-to-end results on CoreSearch development and test sets. **E2E-DPR**: the end-to-end DPR baseline results, where '$^-$' indicates the model was trained without mention boundary tokens, and '$^+$' with them. **E2E-Integrated**: Our end-to-end integrated model.

Reader Evaluation We use the above metrics with the additional question answering (QA) measurements of Exact Match (EM) and token level F1 score[8] with the reference answer after minor normalization as in (Chen et al., 2017; Lee et al., 2019; Karpukhin et al., 2020).[9]

5.3 Results

Retriever Table 4 summarizes the retriever performance results over the CoreSearch test set. Our retriever model surpasses the BM25 method (see further details in Appendix A.1) by a large margin on every metric (Table 4, BM25 versus Retriever-S$^+$). It should be noted that BM25 is considered a strong information retrieval model (Robertson and Zaragoza, 2009), also compared to recent neural-based retrievers (Khattab and Zaharia, 2020; Izacard et al., 2021; Chen et al., 2021; Piktus

[8]Using the official SQuAD evaluation script.

[9]We note that QA measurements only take into consideration lexical matches. However, equal lexical representation does not necessarily imply a coreference relation (for example, the mention *plane crash* can appear twice in the same passage, each time referring to a different plane, thus denoting different events). To that end, we add a necessary constraint limiting relevant answers only to those where the answer index within context intersects with the gold mention span.

No	Query Context	Top-2 Results	Relevancy
1	...Walid al-Maqdisi, a Salafi leader of an al-Qaeda-affiliated terrorist group, responsible for **three bombings** in Dahab in 2006, and which is believed to have close ties with terror cells operating in the Sinai Peninsula...	...to replace it with other measures, such as specific anti-terrorism legislation. The extension was justified by the **Dahab bombings** in April of that year...	✓
		...damaging the industry so that the government would pay more attention to their situation. (See 2004 Sinai bombings, 2005 Sharm El Sheikh bombings and **2006 Dahab bombings**)...	✓
2	...On 14 April 2010, an **earthquake** struck the prefecture, registering a magnitude of 6.9 (USGS, EMSC) or 7.1 (Xinhua). It originated in the Yushu Tibetan Autonomous Prefecture, at local time...	...The airport played an important role in the delivery of rescue personnel and relief supplies to the area affected by the **2010 Yushu earthquake**...	✓
		...China-Congo Friendship Primary School, a school mostly for Tibetan orphans in Chindu County, Qinghai, after the **2010 Yushu earthquake** destroyed the old school...	✓
3	...He made his AFL debut in the 2010 season and was rewarded with an **AFL Rising Star** nomination. He spent six seasons with Essendon, which peaked with a fifth-place finish in the best and fairest, and after 114 games with the club, he was traded to the Melbourne Football Club during the 2015 trade period...	...Aaron Joseph was nominated for the **2009 AFL Rising Star** award for his performance in Carlton's Round 12 win against. Joseph did not poll votes in the final count...	✗
		...Davis made his AFL debut for Adelaide in Round 4, 2010 against Carlton at AAMI Stadium; he had 16 possessions and seven marks. Davis was nominated for the **2010 Rising Star** in round 16...	✓
4	...In the Tang Dynasty, 10 emperors were buried in Weinan after their death. On the morning of 23 January 1556, **the deadliest earthquake on record** with its epicenter in Huaxian killed approximately 830,000 people...	...including the **1556 Shaanxi earthquake** that reportedly killed more than 830,000 people, listed as the deadliest earthquakes of all times and the third deadliest natural disaster...	✓
		...In 1556, during the rule of the Jiajing Emperor, the **Shaanxi earthquake** killed about 830,000 people, the deadliest earthquake of all time...	✓

Table 6: The top-2 query results given by the E2E-Integrated model on a random sample of *Type-2* cluster queries (§3.1). **Blue** signifies the mention span in the query, **green** signifies a correct mention detection, and **purple** signifies a wrong mention detection. The relevancy indicator column signifies whether the retrieved passage itself is relevant or not.

et al., 2021). We observed this phenomenon during our experiments, as the underlying DPR retriever (i.e., BERT without boundary tokens), yielded poor results on our settings, surpassed by the BM25 model on all measurements by a significant gap (Table 4, BM25 versus Retriever-B).

End-to-end Table 5 presents our end-to-end system results applied over the CoreSearch test set. We found that both of the reader models (*E2E-DPR* and *E2E-Integrated*) present appealing performance given different measurement aspects we now describe.

We conclude from the recall results (R@10 and R@50) that the E2E-DPR$^-$ model is an effective re-ranking model, ranking almost all relevant passages extracted by the retriever within the top 50 results (86.62% out of maximum of 87.12% ranked by the Retriever-S$^+$ model at top-500). The EM and F1 results indicate that the E2E-Integrated model gains better mention extraction capabilities compared to both E2E-DPR models (by 1.5% EM and 1.2% F1 compared to E2E-DPR$^+$).

Finally, the MRR and mAP results indicate that the E2E-Integrated model overall performs better then both E2E-DPR models at ranking relevant passages at higher ranks (indicated by MRR@10, mAP@10 and mAP@50 in Table 5). In particular, we find that the MRR@10 results are especially appealing (90.06%), showing the model predominately ranks a relevant passage at the first or second position.

Finally, Table 6 illustrates a sample of the E2E-Integrated top-2 model results, on a sample of queries containing mention spans not including event arguments, randomly sampled from five *Type-2* CoreSearch clusters (§3.1). The table Illustrates the model effectiveness in returning relevant passages and the coreferring mention within them.

False Negative Passages We observed that on rare occasions the model returns a relevant passage (and a coreferring mention) marked as negative in the dataset. We sampled 15 queries and manually validated their top-10 answers. We found that from 58 negative results, only 1 was a false negative, indicating that indeed this phenomenon is rather rare and insignificant. Such false negatives can originate either from the WEC-Eng training set (§2.1), or from our destructing passage generation (§3). Notice that, such false negatives can only have a deflating effect on results.

5.4 Ablation Study

To understand further how different model changes affect the results, we conduct several experiments and discuss our findings below. Table 4 presents the retriever model results and Table 5 presents the reader model results on the development set, for some ablations.

Mention Span Boundaries In both our *retriever* and *reader* experiments, we found that adding the span boundary tokens around the query mention, provides a strong signal to the model. In our retriever experiments, while most of the gain to performance was originated by replacing the BERT model with SpanBERT (Retriever-B and Retriever-S$^-$ in Table 4), applying boundary tokens significantly improved performance further all across the board (Retriever-S$^+$ in Table 4).

However, in our *end-to-end* model experiments, we observed that applying boundary tokens will help the model mostly to improve at span detection, while less so at re-ranking (E2E-DPR$^-$ and E2E-DPR$^+$ in Table 5).

Modeling Coreference with QA Our main motivation for replacing the DPR reader passage selection method (Eq. 5), with a coreference scoring one, was to create a better passage selection mechanism for re-ranking. Indeed, this modeling prove efficient both at re-ranking, as well as at mention detection, as indicated by the E2E-Integrated model results in Table 5.

5.5 Qualitative Error Analysis

To analyze prominent error types made by our E2E-Integrated model we sampled 20 query results that were incorrectly ranked at the first position (Table 7 in Appendix A.2 presents a few of these examples). From those 20 results, 18 were indeed identified as incorrect while 2 results were actually correct, that is, including a mention that does corefer with the query event but was missed in the annotation (a false negative).

We observed two main errors types. The first type involved event argument inconsistencies, identified in 10 out of the 18 erroneous results. In these cases, the model identified an event of the same type as the query event, but with non-matching arguments (see examples 3, 4, 5 and 6 in Table 7). This type of error suggests that there is room for improving the model capability in within- and cross-document argument matching. Some illustrating examples in Table 7 for such argument mismatches include "few days later", "also that year", "the town" (examples 3, 4 and 5, respectively).

The second type of error, identified in 8 out of the 18 erroneous results, corresponded to cases where the two contexts of the query and result passages did not provide sufficient information for determining coreference (see examples 1 and 2 in Table 7). Manually analyzing these 8 cases, we found that in 3 of them the coreference relation could be excluded by examining other event mentions in the coreference cluster to which the query belongs. In 7 cases, it was possible to exclude coreference by consulting external knowledge, specifically Wikipedia, to obtain more information either about the event itself or its arguments. Example 1 in the table illustrates a case where Wikipedia could provide conflicting information about the event location (the city of the *Maxim restaurant* vs. the city of the query event). Example 2 illustrates a case where Wikipedia provided conflicting information about the event time (the time of the first Republican convention in the query vs. the time of the convention discussed in the result). This error type suggests the potential for incorporating external knowledge in cross-document event coreference models. Further, models may benefit from considering globally the information across an entire coreference cluster, as was previously proposed in some works (Raghunathan et al., 2010).

6 Conclusions

We introduced *Cross-document Coreference Search*, a challenging task for accurate semantic search for events. To support research on this task, we created the Wikipedia-based *CoreSearch* dataset, comprised of training, validation, and test set queries, along with a large collection of about 1M passages to retrieve from in each set. Furthermore, our methodology for semi-automatically converting a cross-document event coreference dataset to a coreference search dataset can be applied to other such datasets, for example HyperCoref (Bugert et al., 2021) which represents the news domain. Finally, we provide several effective baseline models and encourage future research on this promising and practically applicable task, hoping that it will lead to a broad set of novel applications and use-cases.

7 Limitations

In this work, we construct the CoreSearch dataset, which relies on the existing Wikipedia Event Coreference dataset (WEC-Eng) (Eirew et al., 2021). This setup exposes potential limitations of the available annotations in WEC-Eng which might be partially noisy in several manners.

By using Wikipedia as the knowledge source, we assume that the corpus is comprised of high quality documents. Yet, future work may further assess the quality of the documents inside WEC-Eng, such as checking for duplications.

Second, since the WEC-Eng train set was built using automatic annotation, it might contain some wrong coreference annotations. Wikipedia instructs authors to mark the first occurrence of a mention in the article. However, for several rare occasions, such distracting passages might contain events which were not covered either due to an author not following the instructions or the existence of more than one mentions of the same event within the same passage (§5.3). While we observed that false-negative retrievals are quite rare, this aspect may be further investigated.

Finally, our dataset covers events which are "famous" to a certain extent, justifying a Wikipedia entry, but does not cover anecdotal events that may arise in various realistic use cases.

Acknowledgments

We thank the Deepset team for providing and supporting the Haystack framework. This research was supported in part by Intel Labs, the Israel Science Foundation grant 2827/21, by a grant from the Israel Ministry of Science and Technology and by the PBC Fellowship for outstanding data science students.

References

Emily Allaway, Shuai Wang, and Miguel Ballesteros. 2021. Sequential cross-document coreference resolution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4659–4671, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Petr Baudiš and Jan Šedivý. 2015. Modeling of the question answering task in the yodaqa system. In *Proceedings of the 6th International Conference on Experimental IR Meets Multilinguality, Multimodality, and Interaction - Volume 9283*, CLEF'15, page 222–228, Berlin, Heidelberg. Springer-Verlag.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on Freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544, Seattle, Washington, USA. Association for Computational Linguistics.

Michael Bugert and Iryna Gurevych. 2021. Event coreference data (almost) for free: Mining hyperlinks from online news. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 471–491, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Michael Bugert, Nils Reimers, and Iryna Gurevych. 2021. Generalizing Cross-Document Event Coreference Resolution Across Multiple Corpora. *Computational Linguistics*, 47(3):575–614.

Avi Caciularu, Arman Cohan, Iz Beltagy, Matthew Peters, Arie Cattan, and Ido Dagan. 2021. CDLM: Cross-document language modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2648–2662, Punta Cana, Dominican Republic. 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 preprint arXiv:2009.11032*.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Xilun Chen, Kushal Lakhotia, Barlas Oğuz, Anchit Gupta, Patrick Lewis, Stan Peshterliev, Yashar Mehdad, Sonal Gupta, and Wen-tau Yih. 2021. Salient phrase aware dense retrieval: Can a dense retriever imitate a sparse one?

Eunsol Choi, Jennimaria Palomaki, Matthew Lamm, Tom Kwiatkowski, Dipanjan Das, and Michael Collins. 2021. Decontextualization: Making Sentences Stand-Alone. *Transactions of the Association for Computational Linguistics*, 9:447–461.

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 Languages Resources Association (ELRA).

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, Zhilin 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.

Alon Eirew, Arie Cattan, and Ido Dagan. 2021. WEC: Deriving a large-scale cross-document event coreference dataset from Wikipedia. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2498–2510, Online. 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.

William Held, Dan Iter, and Dan Jurafsky. 2021. Focus on what matters: Applying discourse coherence theory to cross document coreference. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1406–1417, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 2021. Efficiently teaching an effective dense retriever with balanced topic aware sampling.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Benjamin Hsu and Graham Horwood. 2022. Contrastive representation learning for cross-document coreference resolution of events and entities. *arXiv preprint arXiv:2205.11438*.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning.

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, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

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.

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.

O. Khattab and Matei A. Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the*

57th Annual Meeting of the Association for Computational Linguistics, pages 6086–6096, Florence, Italy. 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 Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. 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 687–692, New Orleans, Louisiana. 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.

Xueling Lin, Haoyang Li, Hao Xin, Zijian Li, and Lei Chen. 2020. KBPearl: A Knowledge Base Population System Supported by Joint Entity and Relation Linking. *Proceedings of the VLDB Endowment*, 13(7):1035–1049.

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*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *ICLR*.

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.

Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Dmytro Okhonko, Samuel Broscheit, Gautier Izacard, Patrick Lewis, Barlas Oğuz, Edouard Grave, Wen-tau Yih, et al. 2021. The web is your oyster–knowledge-intensive nlp against a very large web corpus. *arXiv preprint arXiv:2112.09924*.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. 2010. A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 492–501, Cambridge, MA. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Ellen M. Voorhees. 1999. The trec-8 question answering track report. In *In Proceedings of TREC-8*, pages 77–82.

Piek Vossen, Filip Ilievski, Marten Postma, and Roxane Segers. 2018. Don't annotate, but validate: a data-to-text method for capturing event data. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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. 2020. CorefQA: Coreference resolution as query-based span prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6953–6963, Online. Association for Computational Linguistics.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations*.

Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019. End-to-end open-domain question answering with BERTserini. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 72–77, Minneapolis, Minnesota. 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*.

A Appendices

A.1 Sparse Passage Retriever

We created a BM25 baseline model following common practice of comparing a retriever model with traditional sparse vector space methods such as BM25 (Karpukhin et al., 2020; Khattab and Zaharia, 2020). Additionally, our training procedure depends on challenging negative examples provided by a BM25 model (§4.1).

In our task settings, a query is represented by a context with mention, to that end, we experiment using different query configurations in order to maximize our BM25 results. This included; using the entire query context, the query sentence, decontextualization (Choi et al., 2021) based on the sentence containing the event mention, and using the mention span followed by the Named Entities[10] from the surrounding context. We found the latter to gave us the best BM25 results (*BM25* in Table 4).

A.2 A Sample of Erroneous Top Ranked Results

[10]Using spaCy (Honnibal and Montani, 2017) NER

No	Query Context	Top Result	Error Type
1	On March 4, 2001, while on his way to his office, Dean was critically wounded in a **Palestinian suicide attack** which took place in the centre of Netanya. Dean was rushed to the Hillel Yaffe Medical Center and died five hours later from his wounds. His daughter's sister, Shlomit Ziv, whom he met by chance right before the attack took place, was killed instantly in the attack. Dean was buried in the Tel Mond cemetery. After his death, the first Council house in Tel Mond was named after him - "Naphtali Building"...	Speaking before the United Nations Security Council on 24 June 2017, Israeli ambassador Danny Danon, together with Oran Almog, one of the victims of the **Maxim restaurant suicide bombing**, demanded that the PA cease incentivizing terrorism by paying stipends to terrorists	Cannot be determined (Wikipedia)
2	...However, the nascent Republican Party's **first convention** took place in Philadelphia, and the 1860 elections saw the Republican Party win the state's presidential vote and the governor's office. After the failure of the Crittenden Compromise, the secession of the South, and the Battle of Fort Sumter, the Civil War began with Pennsylvania as a key member of the Union. Despite the Republican victory the 1860 election, Democrats remained powerful in the state, and several "copperheads" called for peace during the war. The Democrats re-took control of the state legislature in the 1862 election, but incumbent Republican Governor Andrew Curtin retained control of the governorship in 1863. In the 1864 election, President Lincoln narrowly defeated Pennsylvania native George B. McClellan for the state's electoral votes	Howe was elected as a Whig to the Thirty-second and Thirty-third Congresses. He was not a candidate for renomination in 1854. He resumed his former business pursuits, and was a delegate to the **1860 Republican National Convention** that nominated Abraham Lincoln as the candidate for president. He was assistant adjutant general on the staff of Governor Andrew Gregg Curtin and chairman of the Allegheny County committee for recruiting Union soldiers during the American Civil War. He was one of the organizers and first president of the Pittsburgh chamber of commerce. He died in Pittsburgh in 1877 and was interred in Allegheny Cemetery	Cannot be determined (Cluster / Wikipedia)
3	A colorless version of the logo is particularly used on a local homepage in recognition of a major tragedy, often for several days. The design was apparently first used on the Google Poland homepage following the air disaster that killed, among others, Polish President Lech Kaczyński in April 2010. A few days later, the logo was used in China and Hong Kong to pay respects to the victims of the Qinghai **earthquake**	She donated her prize money from the tournament and spent time helping the victims and post-reconstruction effort of the **12 May earthquake** that killed nearly 70,000 people and left five to ten million homeless in her home province Sichuan. She did the same with her French Open prize money earlier in the year	Time and location mismatch
4	Swift was named "Billboard"s Woman of the Year in 2014, becoming the first artist to win the award twice. Also that year, she received the Dick Clark Award for Excellence at the **American Music Awards**. In 2015, "Shake It Off" was nominated for three Grammy Awards, including Record of the Year and Song of the Year and Swift won the Brit Award for International Female Solo Artist...	Bieber performed the song on "The Ellen DeGeneres Show" on November 13, 2015. He was also a musical guest on "The Tonight Show Starring Jimmy Fallon". Additionally, Bieber performed the song during the **2015 American Music Awards**, which took place at Microsoft Theater on 22 November 2015 in Los Angeles, California. The singer also took the stage to perform "Sorry"...	Year mismatch
5	Johannes Barge (23 March 1906 – 28 February 2000) was an officer in the Wehrmacht of Nazi Germany during World War II who was responsible for German military operations causing **Cephalonia Massacre** in September 1943	...On 18 June 1944, EDES forces with Allied support launched an attack on Paramythia. After short-term conflict against a combined Cham-German garrison, the town was finally under Allied command. Soon after, violent reprisals were carried out against the town's Muslim community, which was considered responsible for the **massacre of September 1943**	Location mismatch
6	The region had previously experienced one of the worst earthquakes in 1897, measuring 8.1 on the Richter scale, that claimed the lives of over 1,500 people. Again in September 2011, more than 50 people died after a **killer quake measuring 6.9** had shook the region	The 7.2 **Dalbandin earthquake** shook a remote region of Balochistan on 19 January 2011. The dip-slip shock had a maximum Mercalli intensity of VI ("Strong"), caused moderate damage, and left three dead and several injured	Several mismatched entities

Table 7: A sample of queries with top result marked as false (i.e., containing an event not coreferring with the query event), produced by the E2E-Integrated model. **Green** signifies an event mention span. Blue represents some of the event arguments (such as time, location, participants, etc.) that may indicate whether the query and result events hold a coreference relation or not.