Contrastive Representation Learning for Cross-Document Coreference Resolution of Events and Entities

Benjamin Hsu and **Graham Horwood**
AWS AI Labs
{benhsu, ghorwood}@amazon.com

Abstract

Identifying related entities and events within and across documents is fundamental to natural language understanding. We present an approach to entity and event coreference resolution utilizing contrastive representation learning. Earlier state-of-the-art methods have formulated this problem as a binary classification problem and leveraged large transformers in a cross-encoder architecture to achieve their results. For large collections of documents and corresponding set of n mentions, the necessity of performing n^2 transformer computations in these earlier approaches can be computationally intensive. We show that it is possible to reduce this burden by applying contrastive learning techniques that only require n transformer computations at inference time. Our method achieves state-of-the-art results on a number of key metrics on the ECB+ corpus and is competitive on others.

1 Introduction

Coreference resolution is the fundamental NLP task of finding all mentions that refer to the same real world entity or event in text. It is an important step for higher level NLP tasks involving natural language understanding, such as text summarization (Azzam et al., 1999), information extraction (Zelenko et al., 2004), and question-answering (Vicedo and Ferrández, 2000). Historically, coreference resolution of entities in the same text document – within document (WD) coreference resolution – has received the most attention, though more recently focus has moved toward cross-document (CD) coreference resolution.

CD coreference resolution has recently gained renewed interest for its application in multi-document analysis tasks. CD coreference resolution presents unique challenges not found in the WD context. Spans of text come from different documents without any inherent linear order, and there is no notion that antecedents for a given expression typically occur before the expression, as in a single document. Coreferent expressions also cannot be assumed to occur near one another. Furthermore, documents are also assumed to be authored independently and about different—though lexically similar—topics. For instance, the event described in the sentences from Topic 19 in Table 1 below are not coreferential, despite their lexical similarity ("killed").

Another important aspect of CD coreference resolution is the potential scale of the problem. In certain applications, the number of documents can be large and ever growing. In particular, for applications that merge information from across documents, such as multi-document summarization (Falke et al., 2017) or multi-hop question answering (Dhingra et al., 2018), the corpus in question can be both large and dynamically increasing in size.

Past methods of CD coreference resolution have treated the problem as a binary classification task: given two pairs of mentions, classify them as referring to the same entity or not (Bejan and Harabagiu, 2010; Yang et al., 2015; Huang et al., 2019; Kenyon-Dean et al., 2018). In more recent works, contextual embeddings using a cross-encoder architecture have been leveraged to obtain state-of-the-art results (Yu et al., 2020; Zeng et al., 2020; Caciularu et al., 2021) on the ECB+ corpus. Despite achieving state-of-the-art results on the benchmark dataset, a shortcoming of these approaches is the fact they use a transformer as a cross-encoder – two sentences are passed to through the transformer network and a label is predicted. For n mentions in a corpus, these approaches require n^2 comparisions at inference time. As Reimers and Gurevych (2019) noted when using BERT in a cross-encoder architecture, finding the most similar pair of sentences in a collection of $n = 10000$ sentences requires $n(n-1)/2 = 49\,995\,000$ inference computations, which they estimated to take 65 hours using a V100

	Subtopic 1	Subtopic 2
Topic 19	Riots Erupt Following Death of *Brooklyn Teen* **Killed** By *Police*	INITIAL results from the post-mortem on a 15-year-old Greek *boy whose* **killing** by *police* sparked five days of rioting show *Alexandros Grigoropoulos* **died** from a bullet ricochet.
	Yesterday , the *police* explained that officers shot and **killed** a 16-year-old *Kimani Gray* in *Brooklyn* because *he* allegedly pointed a gun at the cops.	Fresh riots were reported in Greece on Saturday December 13 2008 in protest at the **killing** by *police* of a 15-year-old boy, *Alexandros Grigoropoulos*, eight days ago

Table 1: Examples of cross-document coreference clusters from topics 19 of the ECB+ corpus. Bold text indicate events and the same color indicates that they belong in the same coreference cluster. The addition of lexically similar second subtopic (riots in Greece over teenagers death vs riots in Brooklyn over teenagers death) adds an additional challenge to the ECB+ corpus.

GPU.

Others have sought to address the quadratic scaling of these methods. Recently, Allaway et al. (2021); Cattan et al. (2021a) introduced methods that require n transformer passes. In this work, we introduce a method using contrastive learning to generate mention representations that are useful for the coreference resolution problem. Previous attempts along these lines by Kenyon-Dean et al. (2018) introduced clustering-oriented regularization terms in the loss function. Our method improves on these earlier methods on the benchmark dataset, and achieves results competitive with the more expensive methods of Yu et al. (2020); Zeng et al. (2020); Caciularu et al. (2021). We conduct extensive ablations of our model which we discuss in §4.5. We discuss applications to domains outside of the ECB+ corpus in §4.6.

2 Related Work

Most recent work on CD coreference resolution has focused on the ECB+ corpus (Cybulska and Vossen, 2014), which we also use in this work. The ECB+ corpus, which is an extension of the Event Coreference Bank (ECB), consists of documents from Google News clustered into topics and annotated for event coreference (Bejan and Harabagiu, 2010). ECB+ increases the difficulty level of the original ECB dataset by adding a second set of documents for each topic (subtopic), discussing a different event of the same type (e.g. riots in Greece over teenagers death vs riots in Brooklyn over teenagers death; see Table 1) (Cybulska and Vossen, 2014). While relatively small, the corpus is representative of the common cross-document coreference use cases across a restricted set of related documents (i.e. results from a search query).

Most approaches to CD coreference resolution address the problem as a binary classification problem between all pairs of events and entities. Early works utilized hand engineered lexical features (e.g. head lemma, word embedding similarities, etc.) (Bejan and Harabagiu, 2010; Yang et al., 2015). More recent works have relied on neural network methods, utilizing character-based embeddings (Huang et al., 2019; Kenyon-Dean et al., 2018) or contextual embeddings (Yu et al., 2020; Cattan et al., 2020; Zeng et al., 2020; Caciularu et al., 2021; Allaway et al., 2021). Recent approaches by Yu et al. (2020) and Caciularu et al. (2021) leveraging RoBERTa and Longformer transformer models have set strong benchmarks. A drawback of these approaches is the necessity to consider all pairs of n mentions in a corpus in a cross encoder architecture. Each unique pair of entities (separated by a special token) is passed through a transformer to generate a similarity score. This requires n^2 transformer computations.

This can be computationally expensive and several works have sought to address this. Allaway et al. (2021) introduced a model that clusters mentions sequentially at inference time. They achieved competitive results using a BERT-base model and without using a hierarchical clustering algorithm to generate coreference chains. Cattan et al. (2021a) adapted the model of Lee et al. (2017) to the cross-document context. Specifically, they pruned document spans down to the gold mentions and encode each resulting pared document using a RoBERTa-large model. A pairwise (feed-forward network) scorer then generates a score for each pair of spans. They also considered an end-to-end system where they use their model to predict mention spans instead of using gold mentions. In this work, we consider gold mentions only as has been done in earlier works.

In this work, we introduce a method leveraging contrastive learning using a RoBERTa-large model as the base encoder. At inference time, our method requires n passes of the transformer, like earlier methods by Allaway et al. (2021); Cattan et al. (2021a). Our method surpasses their methods on the benchmark ECB+ dataset and is competitive with more expensive cross-encoder approaches of Yu et al. (2020); Zeng et al. (2020); Caciularu et al. (2021).

3 Methodology

3.1 Dataset

We follow earlier works and use the ECB+ corpus, which is an extension of the Event Coreference Bank (ECB), which was discussed in the previous section. Following earlier works by others (Yu et al., 2020; Cattan et al., 2020; Caciularu et al., 2021; Allaway et al., 2021), we follow the setup of Cybulska and Vossen (2015), which was also used by others (Yu et al., 2020; Cattan et al., 2020; Caciularu et al., 2021; Allaway et al., 2021). This setup uses a subset of the annotations which has been validated for correctness and allocates a larger portion of the dataset for training. In this setup, we use topics 1-35 as the train set, setting aside topics 2, 5, 12, 18, 21, 23, 34, 35 for hyperparameter tuning, and 36- 45 as the test set. To preprocess mentions, we utilized the reference implementation from Cattan et al. (2020). The distribution of the train, test, and development sets can be seen in Table 2.

	Train	Dev	Test
# Topics	25	8	10
# Documents	574	196	206
# Event Mentions	3808	1245	1780
# Event Singletons	1116	280	623
# Event Clusters	1527	409	805
# Entity Mentions	4758	1476	2055
# Entity Singletons	814	205	412
# Entity Clusters	1286	330	608

Table 2: Statistics for the ECB+ corpus. We followed the setup of (Cybulska and Vossen, 2015) and used topics 36-45 for our test set and topics 1-35 for training with topics 2, 5, 12, 18, 21, 23, 34, 35 set aside in the development set for hyperparameter tuning.

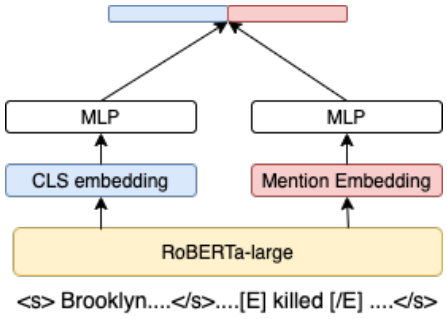

Figure 1: Our encoder takes as input the first two sentences from the document and concatenates it with the sentence containing the mention, taking care to annotate the mention location with tags $[E]$ and $[/E]$.

3.2 Model

We propose a model to learn embeddings useful for clustering events and entities. Our model leverages a Siamese neural network (Bromley et al., 1993) to fine-tune a RoBERTa-large encoder (see Figure 1). We train and evaluate our model using gold mentions as opposed to predicted mentions in order to focus on the cross-document coreference resolution problem. At inference time, our model generates embeddings for the mentions which are then clustered using an agglomerative clustering algorithm as was done previously by Barhom et al. (2019); Yu et al. (2020); Cattan et al. (2020); Caciularu et al. (2021); Zeng et al. (2020). Below we discuss details of our methodology and training procedure.

Document Context Following Caciularu et al. (2021), we use the observation that other parts of the document provide valuable context to the mentions in question. We extract and encode the first two sentences from the document. This takes advantage of the fact that the articles are news articles and in many cases, much of the relevant information is summarized at the beginning of the document. In most cases, these two sentences are the headline and dateline for the article. In cases where the sentence in question is one of the first two sentences, we take the next sentence in the document.

Contextual Embedding In addition to the document context, we also utilize the sentence that the mention appears in and annotate its location in the sentence using $[E]$ and $[/E]$ tokens. The two sequences are concatenated together using a [SEP] token (see Figure 1). In total, we keep 128 word piece tokens and in cases where the combined input exceeds this, we remove tokens from the end of the context before removing tokens from the sentence

containing the mention.

This combined sequence is encoded using a RoBERTa-large model (Liu et al., 2019), as shown in Figure 1. We fine-tune all layers of the RoBERTa-large model. RoBERTa will produce a representation vector for each token of the input sequence. We then sum up element-wise the token-level representations of the mention and use this as the representation of the mention, v_e. Additionally, we utilize the first token of the sequence v_{cls} as the embedding for the entire document context and mention embedding. Each of these contextual embeddings are passed separately through a multi-layer perceptron (MLP). We found that 1024 for the hidden layer dimension for both MLPs worked well in our experiments.

$$v'_e = MLP_1(v_e); \quad v'_{cls} = MLP_2(v_{cls}) \tag{1}$$

The final representation for the mention i and its context document is given by the concatenation of the two vectors output vectors, indicated by $[.;.]$.

$$v_i = [v'_{cls}; v'_e] \tag{2}$$

At inference time, our model takes in the mention and its context (both the head of the document and its sentence) and generates a 2048 dimensional embedding v_i. A clustering algorithm is applied to embeddings to generate coreference clusters. In order to compare our language model with earlier approaches, we follow earlier works and use an agglomerative clustering model. We use the implementation from scikit-learn[1] and cluster mention representations using the cosine distance metric. Representations within an average threshold distance τ are considered to be in the same cluster (i.e. coreferences).

3.3 Training

To train the model, we consider pairs of sentences – positive samples are pairs of sentences where the mentions are coreferential while negative samples are pairs of sentences where the mentions are not coreferential. Pairs of sentences were chosen from *within* gold topics and were constructed by first computing the similarity between sequences. This focuses our model to learn features to distinguish between the two closely related subtopics, one of the key aspects of the ECB+ corpus. We used SBERT (Reimers and Gurevych, 2019) to embed these sequences initially. Positive pairs were created from sequences that were least similar to one another and negative pairs were selected from the set of pairs most similar to one another, both within a particular subtopic and across subtopics (but still within the same topic). Finally, the negative samples were down-sampled by selecting samples whose similarity was greater than the median similarity among all possible positive sample pairs. The resulting distribution for the pairs can be seen in Table 3.

	Events	**Entities**
# of Pairs	19000	27090
# of Positive	2085	4078
# of Negatives	16915	23012
# of Same Subtopic	13694	18847
# of Different Subtopic	5306	8243
Fraction Positive	0.11	0.15
Fraction Same Subtopic	0.72	0.70
Median pos. similarity score	0.62	0.59
Median neg. similarity score	0.80	0.77

Table 3: Statistics for the contrastive pairs generated. Pairs of sentences were chosen from *within* gold topics and were constructed by first computing the similarity between sequences. Negative samples were down-sampled by selecting samples whose similarity was greater than the median similarity among all possible sample pairs.

The model parameters were then trained using a Siamese network architecture (Chopra et al., 2005) where model weights are shared across both branches. For a given pair of sentences $p = (s_1, s_2)$ and label $y = 1, 0$ where $y = 1$ if the pairs are coreferences and $y = 0$ otherwise, each pair of sentences is encoded using our model. The model was trained by minimizing the contrastive loss, ℓ (Hadsell et al., 2006), as implemented by Reimers and Gurevych (2019),

$$\ell = y * d(i,j)^2 + (1-y) * max(0, m - d(i,j))^2 \tag{3}$$

For our purposes, $d(i,j) = 1 - cos(v_i, v_j)$ is the cosine distance, $m > 0$ is a margin, and y is one if the pairs describe coreferent mentions and zero otherwise. Dissimilar pairs contribute to the loss function only if their distance are within m. The loss pushes the embeddings so that positive pairs are closer together in the embedding space and negative pairs are pushed to be more distant

[1] https://scikit-learn.org

than the margin m.

3.4 Hyperparameters

In our experiments, we used the AdamW optimizer without warmup and found that a batch size of 16 worked well. We utilized Ray (Liaw et al., 2018) for hyperparmeter tuning and specifically the Bayesian optimization search algorithm from scikit-optimize.[2] We performed our experiments on a p3dn.24xlarge with 8 V100 Tensor Core GPUs and chose the dropout rate, learning rate, contrastive margin m and clustering threshold τ to optimize the CoNLL F1 score on the development set gold topics. This was done to learn representations that address the lexical ambiguity in the ECB+ corpus topics. Resulting hyperparameters can be found in Table 4.

	Events	Entities
Epochs	100	50
Learning rate	2e-7	2e-7
Batch Size	16	16
Contrastive margin, m	0.40	0.70
Clustering Threshold, τ	0.2	0.2

Table 4: Hyperparameters for our best performing models on events and entities.

4 Results and Discussion

We evaluate our model using four different measures as is common in earlier works. Specifically, we evaluated our model performance using MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), CEAF-e (Luo, 2005), and LEA (Moosavi and Strube, 2016) metrics. We also evaluate our model using the CoNLL F1, the average of the MUC, B^3, and CEAF-e F1 scores. As a baseline, we also show results from a lemma model that takes each span in question and utilizes spaCy[3] to lemmatize each token. Mentions are clustered based on whether their lemmatized tokens are exact matches or not.

Evaluations on ECB+ test corpus are not without controversy, and we discuss these subtleties in detail below. For the reader familiar with these issues, our main results are discussed in §4.2 and §4.3. We also conduct an ablation study with results in §4.5.

[2]https://github.com/scikit-optimize/scikit-optimize
[3]https://spacy.io/

4.1 Evaluation Settings

Many earlier methods leveraged an initial document clustering (Yu et al., 2020; Zeng et al., 2020; Caciularu et al., 2021; Allaway et al., 2021). As observed by Barhom et al. (2019); Upadhyay et al. (2016), clustering the documents as a preprocessing step and performing pairwise classification on mentions within each cluster provides a strong baseline. Barhom et al. (2019) introduced a K-Means algorithm to cluster documents using TF-IDF scores of the unigrams, bigrams and trigrams, where K is chosen by utilizing the Silhouette coefficient method (Rousseeuw, 1987). Models are then applied to mentions within each cluster.

However, this approach has come under criticism (Cremisini and Finlayson, 2020; Cattan et al., 2021a,b). Detractors note that, because of the high lexical similarity between documents within the same subtopic, pre-clustering methods are able to produce near perfectly predicted subtopics, especially in the ECB+ corpus, where only a few coreference links are found across different subtopics. Document clustering is not expected to perform as well in realistic settings where coreferent mentions can spread over multiple topics (Cattan et al., 2021a). More importantly, this bypasses the intention behind the inclusion of subtopics in ECB+ and avoids challenging the coreference models on lexical ambiguity (Cybulska and Vossen, 2014).

In our view, evaluation utilizing the original topic clusters ("gold" topics) is more in line with the original intent of Cybulska and Vossen (2014) and more indicative of realistic settings (Cattan et al., 2021b). We discuss results (1) using ECB+ topics ("gold topics" henceforth) as the initial document clustering and (2) using no initial document clustering ("corpus level" henceforth) in section §4.2. We find that our methodology improves on earlier methods (Tables 5 and 9). Finally, because a majority of earlier works evaluate their models using predicted topics, we discuss our model performance under this setting in §4.3. We report results from a single run.

4.2 Gold Topics and Corpus Level

We evaluate our models using the ECB+ topics, in line with the intent of Cybulska and Vossen (2014) and earlier works by Cattan et al. (2021a,b). According to those authors, this setting was designed to approximate an unclustered stream of news articles.

			MUC			B^3			CEAF-e			LEA			CoNLL
			R	P	F1	R	P	F1	R	P	F1	R	P	F1	F1
Events	Gold Topics	Baseline	72.9	72.4	72.7	51.4	56.5	53.8	58.6	40.4	47.8	46.8	51.5	49.1	58.1
		Cattan et al. (2021b)	80.1	76.3	78.1	63.4	54.1	58.4	56.3	44.2	49.5	59.7	49.6	54.2	62.0
		Ours	**87.8**	**83.0**	**85.3**	**78.0**	**71.4**	**74.5**	**71.0**	**57.4**	**63.5**	**75.6**	**68.9**	**72.1**	**74.4**
	Corpus	Baseline	72.9	60.5	66.1	52	39.2	44.7	48.1	34.7	40.3	46.8	34.6	39.8	50.4
		Cattan et al. (2020)	79.9	74.8	77.2	62.2	48.9	54.8	53.3	42.3	47.2	58.4	44.4	50.5	59.7
		Ours	**86.4**	**74.9**	**80.2**	**76.2**	**51.7**	**61.6**	**59.7**	**48.7**	**53.6**	**73.4**	**48.7**	**58.6**	**65.2**
Entities	Gold Topics	Baseline	61.6	85.9	83.2	31.8	80.2	45.5	53.7	33.8	41.5	28	76.9	41	52.9
		Cattan et al. (2020)	**85.8**	79.3	82.4	64.3	60	62.1	58.6	45.9	51.5	60.9	56.8	58.8	65.3
		Ours	84.5	**90.1**	**87.2**	**72.2**	**80.5**	**76.2**	**73.1**	**59.0**	**65.3**	**69.7**	78.5	**73.8**	**76.2**
	Corpus	Baseline	61.9	77.5	68.8	32.5	67.7	43.9	50.1	33.2	39.9	28.1	63.8	39	50.9
		Cattan et al. (2020)	**85.7**	79.3	82.4	63.7	60	61.8	58.1	45	50.7	60.3	56.8	58.5	65
		Ours	83.9	**86.6**	**85.2**	**71.4**	**75.6**	**73.4**	**69.2**	**55.1**	**61.4**	**68.5**	73.1	**70.7**	**73.3**

Table 5: Combined within- and cross-document coreference scores for entities and events *without* singletons, using gold mentions. Gold topics use the ECB+ topics as the initial document pre-clustering while corpus level results do not use any document pre-clustering. **Bold** values indicate best overall for a particular data subset.

Additionally, as noted by Cattan et al. (2020, 2021a), the presence of singletons biases the results towards models that perform well on detecting all the mentions instead of predicting coreference clusters. Furthermore, in using gold mentions in the evaluation (like we do here), including singletons artificially inflates performance metrics (Cattan et al., 2021a). We present our results without singletons (Table 5) using the reference implementation of Moosavi and Strube (2016). In Appendix A, we give results *with singletons* in Table 9.

On the gold topic and corpus level subsets, our model performs well. In all cases, we surpass the current state-of-the-art model on the CoNLL F1 metric for both event and entity coreference resolution by large margins without singletons (see Table 5)). We suspect this improvement to be a feature of contrastive learning and methodology we used to choose pairs – coreferential mentions are pushed closer together in the embedding space while mentions that are not coreferences are pushed further apart. We do observe a larger drop in performance in going from gold topics to the corpus level subsets. This is due to the choice in contrastive pairs, where negative examples come from the same gold topic.

Aside from improved performance, our methodology differs in some key aspects to the recent works by Cattan et al. (2021a,b, 2020). Their methodology also leverages a RoBERTa-large model to embed documents, but breaks long documents into 512 word piece token chunks. The authors used as their feature vector for a span in question: the sum of the span embeddings, the embeddings for the span beginning and end, and a vector encoding the span length as their feature vector, which they feed into a pairwise classifier to generate pairwise scores. We on the other hand use the sentences containing the span in question and additional context sentences from the document, keeping a total of 128 word piece tokens. This additional context from the document, despite keeping fewer tokens, accounts for much of the performance gain. This is discussed in further detail in §4.5.

4.3 Predicted Topic Clusters

We compare our model against the majority of earlier works that used predicted topic clusters and gold mentions (see Table 6 and Appendix A Table 8 for more complete results). We used the reference implementation by Pradhan et al. (2014) to score our models *with singletons*. Our model is competitive with earlier approaches (Yu et al., 2020; Zeng et al., 2020; Caciularu et al., 2021), despite using significantly fewer resources at inference time – n transformer computations at inference time as oppose to n^2 transformer computations. We also note that in contrast to our approach, Caciularu et al. (2021) used a total of 600 tokens from each document (most documents are within 512 tokens) whereas we only use 128 tokens. Models by Yu et al. (2020); Zeng et al. (2020) employ a BERT based semantic role labelling (SRL) model. On average, our model lags their models by approximately 1.1 CoNLL F1 points, however, we note that Yu et al. (2020) find that the SRL tagging accounted for roughly 0.4 CoNLL F1 points.

When comparing to other models that are linear in transformer computations, our model does well. Compared to the work by Allaway et al. (2021), our model surpasses their results by 3.7 CoNLL

	Scaling	Adapt.	Fine-tuned	SRL	Encoder	System	MUC F1	B^3 F1	CEAF-e F1	CoNLL F1
Events						Baseline	76.7	77.5	73.2	75.7
	n^2		✓	✓	BERT-large	Zeng et al. (2020)	87.5	83.2	**82.3**	84.3
			✓	✓	RoBERTa-large	Yu et al. (2020)	86.6	85.4	81.3	84.4
		✓	✓		Longformer	Caciularu et al. (2021)	**88.1**	**86.4**	82.2	**85.6**
					RoBERTa-large	Cattan et al. (2021a)	83.5	82.4	77.0	81.0
		✓		✓	BERT-base	Allaway et al. (2021)	82.2	81.1	79.1	80.8
			✓			Ours				
	n					– RoBERTa-large	85.6	84.8	79.6	83.3
						– RoBERTa-base	84.0	82.4	79.0	81.8
						– BERT-large	82.8	82.3	77.9	81.0
						– BERT-base	79.8	79.4	74.4	77.9
Entities						Baseline	70.7	61.7	56.9	63.1
	n^2	✓	✓		Longformer	Caciularu et al. (2021)	**89.9**	**82.1**	**76.8**	**82.9**
					RoBERTa-large	Cattan et al. (2021a)	83.6	72.7	63.1	73.1
		✓		✓	BERT-base	Allaway et al. (2021)	84.3	72.4	69.2	75.3
	n		✓		Ours					
						– RoBERTa-large	87.1	80.3	73.1	80.2
						– RoBERTa-base	83.6	74.1	68.5	75.4
						– BERT-large	80.8	71.4	66.2	72.8
						– BERT-base	78.2	68.9	62.7	69.9

Table 6: A comparison of methods utilizing contextual embedding models and their performance on the ECB+ test corpus using *predicted* topic clusters of Barhom et al. (2019). We have indicated the scaling at inference time (in terms of transformer computations) above. We have also indicated whether systems utilized adaptive pre-training (Adapt.), fine-tuned encoders (Fine-tuned), or utilized a semantic role labelling model (SRL). To better compare to earlier works, we have included results from using different encoders in our model and indicated which encoders were used in earlier works. Finally, Allaway et al. (2021) used sequential clustering algorithm whereas ours and Cattan et al. (2020) utilized an agglomerative clustering algorithm. **Bold** indicates best overall. Underlined results indicate our best overall.

F1 points on average. We note however, that their model used a BERT-base model and that they also introduced a novel sequential clustering approach. Our methodology used the larger RoBERTa-large model, and we utilized an agglomerative clustering algorithm as in previous works.

Finally, in contrast to earlier works, we note that our model performs equally well when using predicted clusters and ECB+ gold topics. In fact, our model does better (by 0.9 CoNLL F1 points) on entities when going to gold topics, and achieves the same performance on events using gold topics. This is related to how we selected our contrastive pairs – negative and positive pairs were selected from within each topic and so our model focused on the lexical ambiguity in the ECB+ corpus.

		Entities		Events	
		F1	Δ	F1	Δ
Pred. Topics	Our Model	80.2		83.3	
	– `CLS` representation	77.8	-2.4	82.3	-1.0
	– mention representation	77.8	-2.4	82.0	-1.3
	– no document context	74.2	-6.0	80.8	-2.5
Gold Topics	Our Model	81.1		83.3	
	– `CLS` representation	79.0	-2.1	81.3	-2.0
	- mention representation	78.9	-2.2	79.6	-3.7
	– no document context	75.1	-6.0	77.1	-6.2
Corpus	Our Model	78.7		75.9	
	– `CLS` representation	77.3	-1.4	74.9	-1.0
	– mention representation	75.7	-3.0	72.0	-3.9
	– no document context	72.8	-6.0	70.5	-5.4

Table 7: Ablation results (CoNLL F1) on the ECB+ test set *with singletons*.

4.4 Training and Inference Time

Our model is larger than earlier models by Cattan et al. (2021a,b); Allaway et al. (2021). On a single V100 Tensor Core GPU with 32 GB of RAM, training took approximately two days. This is comparable to reported times for the cross-encoder model (using Longformer) by (Caciularu et al., 2021). We note that contrastive learning methods have been found to converge slowly (Sohn, 2016). At inference time, however, our model takes approximately 15 seconds to evaluate on the ECB+ test set of events (using gold mentions and with singletons included). As a point of comparison, we ran the model of Cattan et al. (2021a,b) which likewise uses RoBERTa-large and is linear in transformer computations. We found that their model takes approximately 60 seconds under similar settings. In §4.5 we discuss experiments with smaller models.

4.5 Ablations

We ablate several parts of our model using the headlines heuristic and examine the importance of the underlying language model, the token representations, and the document context.

Language Model We examined the effect different representations have on overall performance by ablating the language model used. We found that the larger and richer representations of the RoBERTa-large model performed better generically. We gained on average 5 CoNLL F1 points in using RoBERTa-large versus BERT-large. We gained on average 7.2 CoNLL F1 points versus the smaller BERT-base model. Details can be found in Table 6.

Token Representation To assess the effect of including the `CLS` token embedding in the final representations, we trained our model without using its representation, but keeping the mention representation. We find that the `CLS` representation accounts for roughly 1.3 CoNLL F1 points on average while the mention representation accounts for roughly 2.8 CoNLL F1 points on average (see Table 7 for details). We also examined our model without explicitly using the mention representation, but still tagging the span with $[E], [/E]$ tokens. For our model, we find that the mention representation was a more important factor when considering events. We speculate that tagging the mention location with $[E], [/E]$ tokens allows the transformer to attend to the mention. For events, which have a more complicated structure (e.g. arguments) this likely has a more important effect.

Document Context Finally, an important component of our model was including the first two sentences of each document in the spirit of Caciularu et al. (2021). For the ECB+ corpus, which is comprised of news articles, much contextual information is contained in the first two sentences of the document. We see that the document context contributes on average 5.4 CoNLL F1 points (see Table 7 for details). This is in line with our expectations for new articles and earlier observations by Caciularu et al. (2021). We suspect the importance of this feature is due to a property of the ECB+ corpus that has been highlighted by others – namely, the documents form fairly distinct clusters in themselves and so simple document embeddings are able to recover subtopics easily (Cattan et al., 2021a,b; Cremisini and Finlayson, 2020). Note for instance that our model without using document context is competitive (compare with Table 6) when using predicted topics for pre-clustering. Recently, Eirew et al. (2021) sought to address this issue by creating a the Wikipedia Events Coreference (WEC) dataset. Applying our model to the WEC dataset, we found that our results surpass their benchmark by large margins (CoNLL F1 of 89.3 versus 62.3 (Eirew et al., 2021)) . We plan to discuss these results in further detail in future work.

4.6 TextRank

A limitation of the current work is its specificity to formal text (i.e. news articles, Wikipedia articles). Given the importance of the headlines to our model, we also conducted experiments using the TextRank algorithm (Mihalcea and Tarau, 2004) to extract sentences that best summarize the content of the article instead of using the first two. We expect this method to be more applicable to less formal settings. We embedded each sentence in the document using SBERT and select the top two. On average we found that the headlines heuristic provided a 4.6 and 3.7 CoNLL F1 gain on on event and entity coreference resolution respectively (with singletons) over the TextRank extracted contexts (for detailed metrics see Table 8 in Appendix A). This is expected in the ECB+ context as the TextRank algorithm selects noisier sentences as compared to article headlines.

5 Conclusions

In this paper, we proposed a new model for within- and cross-document coreference resolution. We demonstrated that contrastive learning approaches are effective at learning representations for coreference resolution. We evaluated our model on gold topics and at the corpus level of the ECB+ corpus—with and without singleton mentions—and found that our approach surpasses current state-of-the-art methods by large margins. We also evaluated our models with an initial document clustering method and found that our model was competitive with earlier works. We presented extensive ablations of our model and discussed limitations of our work including model size, training time, application to formal text domains (i.e. news articles and Wikipedia), and use of agglomerative clustering to generate final coreference clusters. Interesting directions for future work would be testing the TextRank algorithm

in less formal contexts (i.e. beyond news articles and Wikipedia articles), investigating higher-order tuples (e.g. triplets) to speed up model convergence, and extending our work to predicted mentions as opposed to gold mentions as has been done by others (Cattan et al., 2021a,b).

Acknowledgements

The authors thank the anonymous reviewers for their advice and comments.

Ethical Considerations

In this work, we used the ECB+ corpus (Cybulska and Vossen, 2014) which consists of news articles from the open domain. Our use was consistent with the intended use of the dataset. Our model does not contain any intentional biases. As discussed in §3.4, §4.4, we ran our experiments on a single p3dn.24xlarge with 8 V100 32GB GPUs. Model training and inference was relatively short and does not present ethical issues.

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.

Saliha Azzam, Kevin Humphreys, and Robert Gaizauskas. 1999. Using coreference chains for text summarization. In *Coreference and Its Applications*.

A. Bagga and B. 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.

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

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

Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. 1993. Signature verification using a "siamese" time delay neural network. In *Proceedings of the 6th International Conference on Neural Information Processing Systems*, NIPS'93, page 737–744, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

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.

Arie Cattan, Alon Eirew, Gabriel Stanovsky, Mandar Joshi, and Ido Dagan. 2021a. Cross-document coreference resolution over predicted mentions. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5100–5107, Online. Association for Computational Linguistics.

Arie Cattan, Alon Eirew, Gabriel Stanovsky, Mandar Joshi, and Ido Dagan. 2021b. Realistic evaluation principles for cross-document coreference resolution. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 143–151, Online. Association for Computational Linguistics.

S. Chopra, R. Hadsell, and Y. LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546 vol. 1.

Andres Cremisini and Mark Finlayson. 2020. New insights into cross-document event coreference: Systematic comparison and a simplified approach. In *Proceedings of the First Joint Workshop on Narrative Understanding, Storylines, and Events*, pages 1–10, Online. Association for Computational Linguistics.

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

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

R. Hadsell, S. Chopra, and Y. LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742.

Yin Jou Huang, Jing Lu, Sadao Kurohashi, and Vincent Ng. 2019. Improving event coreference resolution by learning argument compatibility from unlabeled data. 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 785–795, Minneapolis, Minnesota. 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.

Richard Liaw, Eric Liang, Robert Nishihara, Philipp Moritz, Joseph E Gonzalez, and Ion Stoica. 2018. Tune: A research platform for distributed model selection and training. *arXiv preprint arXiv:1807.05118*.

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.

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

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain. 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.

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

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Kihyuk Sohn. 2016. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

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

José L. Vicedo and Antonio Ferrández. 2000. Importance of pronominal anaphora resolution in question answering systems. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 555–562, Hong Kong. Association for Computational Linguistics.

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*, MUC6 '95, page 45–52, USA. Association for Computational Linguistics.

Bishan Yang, Claire Cardie, and Peter Frazier. 2015. A hierarchical distance-dependent bayesian model for event coreference resolution. *Transactions of the Association for Computational Linguistics*, 3(0):517–528.

Xiaodong Yu, Wenpeng Yin, and Dan Roth. 2020. Paired representation learning for event and entity coreference.

Dmitry Zelenko, Chinatsu Aone, and Jason Tibbetts. 2004. Coreference resolution for information extraction. In *Proceedings of the Conference on Reference Resolution and Its Applications*, pages 24–31, Barcelona, Spain. Association for Computational Linguistics.

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.

A Detailed Metrics

Below we give detailed metrics *with singletons*.

		MUC			B^3			CEAF-e			LEA			CoNLL
		R	P	F1	R	P	F1	R	P	F1	R	P	F1	F1
Events	Baseline	72.5	81.1	76.6	69.6	87.4	77.5	77.9	69	73.2	55.63	72.9	63.1	75.7
	Zeng et al. (2020)	85.6	**89.3**	87.5	77.6	89.7	83.2	84.5	80.1	**82.3**	-	-	-	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
	Caciularu et al. (2021)	87.1	89.2	**88.1**	84.9	**87.9**	**86.4**	**83.3**	81.2	82.2	**76.7**	**77.2**	**76.9**	**85.6**
	Cattan et al. (2021a)	85.1	81.9	83.5	82.1	82.7	82.4	75.2	78.9	77	68.8	72	70.4	81
	Allaway et al. (2021)	81.7	82.8	82.2	80.8	81.5	81.1	79.8	78.4	79.1	-	-	-	80.8
	Ours													
	– RoBERTa-large	87.9	83.4	85.6	**86.2**	83.4	84.8	76.9	82.4	79.6	74.1	74.2	74.1	83.3
	– RoBERTa-base	83.6	84.5	84.0	78.9	86.1	82.4	79.5	78.5	79.0	67.1	75.8	71.2	81.8
	– BERT-large	82.9	82.7	82.8	81.3	83.4	82.3	77.8	78.0	77.9	68.9	72.5	70.6	81.0
	– BERT-base	80.3	79.3	79.8	78.0	80.9	79.4	73.8	75.0	74.4	63.4	68.8	66.0	77.9
	– RoBERTa-large + TextRank	80.0	83.6	81.8	76.9	86.4	81.4	78.6	74.7	76.6	64.1	74.3	68.8	79.9
Entities	Baseline	58.7	88.6	70.7	46.2	93.1	61.7	79.7	44.2	56.9	35.6	68.2	46.8	63.1
	Caciularu et al. (2021)	**88.1**	**91.8**	**89.9**	**82.5**	81.7	**82.1**	**81.2**	**72.9**	**76.8**	**76.4**	73	**74.7**	**82.9**
	Cattan et al. (2021a)	85.7	81.7	83.6	70.7	74.8	72.7	59.3	67.4	63.1	56.8	65.8	61	73.1
	Allaway et al. (2021)	83.9	84.7	84.3	74.5	70.5	72.4	70	68.1	69.2	-	-	-	75.3
	Ours													
	– RoBERTa-large	83.1	91.6	87.1	72.2	90.4	80.3	81.1	66.5	73.1	63.7	**79.3**	70.6	80.2
	– RoBERTa-base	77.2	91.1	83.6	61.6	92.8	74.1	81.0	59.4	68.5	52.2	79.2	62.9	75.4
	– BERT-large	72.8	90.7	80.8	58.1	92.7	71.4	81.8	55.6	66.2	49.0	76.6	59.7	72.8
	– BERT-base	69.9	’ 88.7	78.2	55.5	90.9	68.9	78.5	52.2	62.7	45.0	72.3	55.5	69.9
	– RoBERTa-large + TextRank	75.6	91.2	82.7	59.1	**93.2**	72.3	80.8	57.4	67.1	49.6	78.9	60.9	74.1

Table 8: Detailed results comparing methods utilizing contextual embedding models and their performance on the ECB+ test corpus using *predicted* topic clusters. Note that the systems of Zeng et al. (2020); Yu et al. (2020); Caciularu et al. (2021) require significantly more resources than the others (n^2 versus n transformer computations). Finally, Allaway et al. (2021) uses a BERT-base model and a sequential clustering algorithm whereas ours and Cattan et al. (2020) utilize RoBERTa-large models and an agglomerative clustering algorithm.

			MUC			B^3			CEAF-e			LEA			CoNLL
			R	P	F1	R	P	F1	R	P	F1	R	P	F1	F1
Events	Gold Topics	Baseline	72.9	72.4	72.7	69.7	73.5	71.5	71.1	71.7	71.4	53.5	59.2	56.1	71.9
		Cattan et al. (2021b)	80.1	76.3	78.1	77.4	71.7	74.5	73.1	77.8	75.4	62.9	59.1	61	76
		Ours	**87.8**	**82.9**	**85.3**	**86.5**	**83.1**	**84.8**	76.9	**82.8**	**79.7**	**74.4**	**74.0**	**74.2**	**83.3**
	Corpus	Baseline	72.9	60.5	66.1	69.7	56.4	62.4	51.5	68.6	58.8	45.3	42.6	43.9	62.4
		Kenyon-Dean et al. (2018)†	67	71	69	71	67	69	**71**	67	69	-	-	-	69
		Ours	**86.4**	**74.9**	**80.2**	**85.3**	**67.9**	**75.6**	65.3	**80.1**	**71.9**	**68.3**	**57.5**	**62.4**	**75.9**
Entities	Gold Topics	Baseline	61.6	85.9	71.8	48.6	**89**	62.9	76.7	45.9	57.4	37.3	65.5	47.5	64
		Cattan et al. (2021a)	-	-	-	-	-	-	-	-	-	-	-	-	70.9
		Ours	**84.5**	**90.1**	**87.2**	**79.3**	86.6	**82.8**	**78.7**	**68.6**	**73.3**	**70.3**	**75.7**	**72.9**	**81.1**
	Corpus	Baseline	61.9	77.5	68.8	48.7	79.6	60.4	68.2	46.1	55	35.2	57.8	43.7	61.4
		Ours	**83.9**	**86.6**	**85.2**	**78.5**	**82.7**	**80.5**	73.0	**67.9**	**70.4**	**67.8**	**71.7**	**69.7**	**78.7**

Table 9: Combined within- and cross-document coreference scores for entities and events *with* singletons, using gold mentions. Gold topics use the ECB+ topics as the initial document pre-clustering while corpus level results do not use any document pre-clustering. We note that the system proposed by Kenyon-Dean et al. (2018) does not use contextual embeddings whereas ours and Cattan et al. (2021a) make use of RoBERTa-large. To the best of our knowledge, we have the only results at the corpus level for entities. **Bold** values indicate best overall for a particular data subset.