CD^{2}CR: Co-reference Resolution Across Documents and Domains

James Ravenscroft[1,2,3]   **Arie Cattan**[4]   **Amanda Clare**[5]   **Ido Dagan**[4]   **Maria Liakata**[1,2]

[1]: University of Warwick, United Kingdom
[2]: Alan Turing Institute, United Kingdom
[3]: Filament AI, United Kingdom
[4]: Bar-Ilan University, Israel
[5]: Aberystwyth University, United Kingdom

{j.ravenscroft|m.liakata}@warwick.ac.uk
{arie.cattan|ido.k.dagan}@gmail.com
afc@aber.ac.uk

Abstract

Cross-document co-reference resolution (CDCR) is the task of identifying and linking mentions to entities and concepts across many text documents. Current state-of-the-art models for this task assume that all documents are of the same type (e.g. news articles) or fall under the same theme. However, it is also desirable to perform CDCR across different domains (type or theme). A particular use case we focus on in this paper is the resolution of entities mentioned across scientific work and newspaper articles that discuss them. Identifying the same entities and corresponding concepts in both scientific articles and news can help scientists understand how their work is represented in mainstream media. We propose a new task and English language dataset for cross-document cross-domain co-reference resolution (CD^{2}CR). The task aims to identify links between entities across heterogeneous document types. We show that in this cross-domain, cross-document setting, existing CDCR models do not perform well and we provide a baseline model that outperforms current state-of-the-art CDCR models on CD^{2}CR. Our data set, annotation tool and guidelines as well as our model for cross-document cross-domain co-reference are all supplied as open access open source resources.

1 Introduction

Cross-document co-reference resolution (CDCR) is the task of recognising when multiple documents mention and refer to the same real-world entity or concept. CDCR is a useful NLP process that has many downstream applications. For example, CDCR carried out on separate news articles that refer to the same politician can facilitate inter-document sentence alignment required for stance detection and natural language inference

models. Furthermore, CDCR can improve information retrieval and multi-document summarisation by grouping documents based on the entities that are mentioned within them.

Recent CDCR work (Dutta and Weikum, 2015; Barhom et al., 2019; Cattan et al., 2020) has primarily focused on resolution of entity mentions across news articles. Despite differences in tone and political alignment, most news articles are relatively similar in terms of grammatical and lexical structure. Work based on modern transformer networks such as BERT (Devlin et al., 2019) and ElMo (Peters et al., 2018) have been pre-trained on large news corpora and are therefore well suited to news-based CDCR (Barhom et al., 2019).

However, there are cases where CDCR across documents from different domains (i.e. that differ much more significantly in style, vocabulary and structure) is useful. One such example is the task of resolving references to concepts across scientific papers and related news articles. This can help scientists understand how their work is being presented to the public by mainstream media or facilitate fact checking of journalists' work (Wadden et al., 2020). A chatbot or recommender that is able to resolve references to current affairs in both news articles and user input could be more effective at suggesting topics that interest the user. Finally, it may be helpful for e-commerce companies to know when product reviews gathered from third party websites refer to one of their own listings. The work we present here focuses on the first cross-document, cross-domain co-reference-resolution (CD^{2}CR) use case, namely co-reference resolution between news articles and scientific papers.

The objective of CD^{2}CR is to identify co-referring entities from documents belonging to different domains. In this case co-reference resolution is made more challenging by the differences in language use (lexical but also syntactic) across the dif-

270

ferent domains. Specifically, authors of scientific papers aim to communicate novel scientific work in an accurate and unambiguous way by using precise scientific terminology. Whilst scientific journalists also aim to accurately communicate novel scientific work, their work is primarily funded by newspaper sales and thus they also aim to captivate as large an audience as possible. Therefore journalists tend to use simplified vocabulary and structure, creative and unexpected writing style, slang, simile, metaphor and exaggeration to make their work accessible, informative and entertaining in order to maximise readership (Louis and Nenkova, 2013).

Success at the CD^2CR task in this setting is dependent on context sensitive understanding of how the accessible but imprecise writing of journalists maps on to precise terminology used in scientific writing. For example, a recent study has found that "convalescent plasma derived from donors who have recovered from COVID-19 can be used to treat patients sick with the disease" [1]. A news article[2] discussing this work says that "...blood from recovered Covid-19 patients in the hope that transfusions...[can help to treat severely ill patients]" . In this example the task is to link 'blood' to 'convalescent plasma' and 'recovered Covid-19 patients' to 'donors'. These cross-document, cross-domain co-reference chains can be used as contextual anchors for downstream analysis of the two document settings via tasks such as natural language inference, stance detection and frame analysis.

The contributions in this paper are the following:

- A novel task setting for CDCR that is more challenging than those that already exist due to linguistic variation between different domains and document types (we call this CD^2CR).

- An open source English language CD^2CR dataset with 7602 co-reference pair annotations over 528 documents and detailed 11 page annotation guidelines (section 3.1).

- A novel annotation tool to support ongoing data collection and annotation for CD^2CR including a novel sampling mechanism for calculating inter-annotator agreement (Section 3.4).

- A series of experiments on our dataset using different baseline models and an indepth capability-based evaluation of the best-performing baseline (Section 5)

2 Related Work

2.1 Co-reference Resolution

Intra-document co-reference resolution is a well understood task with mature training data sets (Weischedel et al., 2013) and academic tasks (Recasens et al., 2010). The current state of the art model by Joshi et al. (2020) is based on Lee et al. (2017, 2018) and uses a modern BERT-based (Devlin et al., 2019) architecture. Comparatively, CDCR, which involves co-reference resolution across multiple documents, has received less attention in recent years (Bagga and Baldwin, 1998; Rao et al., 2010; Dutta and Weikum, 2015; Barhom et al., 2019). Cattan et al. (2020) jointly learns both entity and event co-reference tasks, achieving current state of the art performance for CDCR, and as such provides a strong baseline for experiments in CD^2CR. Both Cattan et al. (2020) and Barhom et al. (2019) models are trained and evaluated using the ECB+ corpus (Cybulska and Vossen, 2014) which contains news articles annotated with both entity and event mentions.

2.2 Entity Linking

Entity Linking (EL) focuses on alignment of mentions in documents to resources in an external knowledge resource (Ji et al., 2010) such as SNOMED CT[3] or DBPedia[4]. EL is challenging due to the large number of pairwise comparisons between document mentions and knowledge resource entities that may need to be carried out. Raiman and Raiman (2018) provide state of the art performance by building on Ling et al. (2015)'s work in which an entity type system is used to limit the number of required pairwise comparisons to related types. Yin et al. (2019) achieved comparable results using a graph-traversal method to similarly constrain the problem space to candidates within a similar graph neighbourhood. EL can be considered a narrow sub-task of CDCR since it cannot resolve novel and rare entities or pronouns (Shen et al., 2015). Moreover EL's dependency on expensive-to-maintain external knowledge graphs is also problematic when limited human expertise is available.

Given these limitations, EL is inappropriate within our task setting, hence our CDCR-based approach.

2.3 Semantic Specialisation

Like earlier static vector language models, contextual language models such as BERT (Devlin et al., 2019) and ElMo (Peters et al., 2018) use distributional knowledge (Harris, 1954) inherent in large text corpora to learn context-aware word embeddings that can be used for downstream NLP tasks. However, these models do not learn about formal lexical constraints, often conflating different types of semantic relatedness (Ponti et al., 2018; Lauscher et al., 2020). This is a weakness of all distributional language models that is particularly problematic in the context of CD^{2}CR for entity mentions that are related but not co-referent (e.g. "Mars" and "Jupiter") as shown in section 5.

A number of solutions have been proposed for adding lexical knowledge to static word embeddings (Yu and Dredze, 2014; Wieting et al., 2015; Ponti et al., 2018) but contextual language models have received comparatively less attention. Lauscher et al (2020) propose adding a lexical relation classification step to BERT's language model pre-training phase to allow the model to integrate both lexical and distributional knowledge. Their model, LIBERT, has been shown to facilitate statistically-significant performance boosts on a variety of downstream NLP tasks.

3 Dataset creation

Our dataset is composed of pairs of news articles and scientific papers gathered automatically (Section 3.1). Our annotation process begins by obtaining summaries of the news and science document pairs (extractive news summaries and scientific abstracts, respectively) (Section 3.2). Candidate co-reference pairs from each summary-abstract pair are identified and scored automatically. (Section 3.3). Candidate co-reference pairs are then presented to human annotators via a bespoke annotation interface for scoring (Section 3.4). Annotation quality is measured on an ongoing basis as new candidates are added to the system (Section 3.5).

3.1 Data Collection

We have developed a novel data set that allows us to train and evaluate a CD^{2}CR model. The corpus is approximately 50% the size of the ECB+ corpus (918 documents) (Cybulska and Vossen, 2014)

Subset	Documents	Mentions	Clusters
Train	300	4,604	426
Dev	142	1,821	199
Test	86	1,177	101

Table 1: Total individual documents, mentions, co-reference clusters of each subset excluding singletons.

and is split into training, development and test sets (statistics for each subset are provided in Table 1). Each pair of documents consists of a scientific paper and a newspaper article that discusses the scientific work. In order to detect pairs of documents, we follow the approach of (Ravenscroft et al., 2018), using approximate matching of author name and affiliation metadata, date of publishing and exact DOI matching where available to connect news articles to scientific publications.

We built a web scraper that scans for new articles from the 'Science' and 'Technology' sections of 3 well-known online news outlets (BBC, The Guardian, New York Times) and press releases from Eurekalert, a widely popular scientific press release aggregator. Once a newspaper article and related scientific paper are detected, the full text from the news article and the scientific paper abstract and metadata are stored. Where available the full scientific paper content is also collected. We ran the scraper between April and June 2020 collecting news articles and scientific papers including preprints discussing a range of topics such as astronomy, computer science and biology (incl. coverage of COVID-19). New relevant content is downloaded and ingested into our annotation tool (see Section 3.4) on an ongoing basis as it becomes available.

3.2 Article Summarisation

Newspaper articles and scientific papers are long and often complex documents, usually spanning multiple pages, particularly the latter. Moreover the two document types differ significantly in length. Comparing documents of such uneven length is a difficult task for human annotators. We also assume that asking human annotators to read the documents in their entirety to identify co-references would be particularly hard with a very low chance for good inter-annotator agreement (IAA). We therefore decided to simplify the task by asking annotators to compare summaries of the newspaper article (5-10 sentences long) and the scientific paper (abstract).

For each document pair, we ask the annotators to identify co-referent mentions between the scientific paper abstract and a summary of the news article that is of similar length (e.g. 5-10 sentences). Scientific paper abstracts act as a natural summary of a scientific work and have been used as a strong baseline or even a gold-standard in scientific summarisation tasks (Liakata et al., 2013). Furthermore, abstracts are almost always available rather than behind paywalls like full text articles. For news summarisation, we used a state-of-the-art extractive model (Grenander et al., 2019) to extract sentences forming a summary of the original text. This model provides a summary de-biasing mechanism preventing it from focusing on specific parts of the full article, preserving the summary's informational authenticity as much as possible.

The difference in style between the two documents is preserved by both types of summary since abstracts are written in the same scientific style as full papers and the extractive summaries use verbatim excerpts of the original news articles.

3.3 Generation of pairs for annotation

To populate our annotation tool, we generate pairs of candidate cross-document mentions to be evaluated by the user. Candidate mentions are identified by using spaCy (Honnibal and Montani, 2017) for the recognition of noun phrases and named entities from each input document pair (abstract-news summary). For each pair of documents, pairs of all possible mention combinations are generated and stored for annotation.

In any given pair of documents, the majority of mention pairs (M_0, M_1) generated automatically in this way will not co-refer thus resulting in a vastly imbalanced dataset and also running the risk of de-motivating annotators. To ensure that annotators are exposed to both positive and negative examples, we use a similarity score to rank examples based on how likely they are to co-refer. The first step in generating a similarity score s is to concatenate each abstract-news-summary pair together: "summary [SEP] abstract" into a pre-trained $BERT_{large}$ model. Then we take the mean of the word vectors that correspond to the mention spans within the documents and calculate the cosine similarity of these vectors. We find that this BERT-based similarity score performs well in practice. We also use it in combination with a thresholding policy as one of our baseline models in Section 4.

3.4 Annotation Tool & Interface

We developed an open source annotation tool[5] that allows humans to identify cross-document co-reference between each pair of related documents. Whilst designing this tool, we made a number of decisions to simplify the task and provide clear instructions for the human annotators in order to encourage consistent annotation behaviour.

To maximise the quality and consistency of annotations in our corpus, we simplified the task as much as possible for the end user. Annotation tasks were framed as a single yes or no question: "Are **x** and **y** mentions of the same entity?". Mentions in context were shown in bold font whereas mentions already flagged as co-referent were shown in green. This enabled annotators to understand the implications for existing co-reference chains before responding (see Figure 3). Questions were generated and ranked via our task generation pipeline (see Section 3.3 above).

We added two additional features to our annotation interface to improve annotators' experience and to speed up the annotation process. Firstly, if the candidate pair is marked as co-referent, the user is allowed to add more mentions to the coreference cluster at once. Secondly, inspired by (Li et al., 2020), if the automatically shown mention pair is not co-referent, the user can select a different mention that is co-referent.

The upstream automated mention detection mechanism can sometimes introduce incomplete or erroneous mentions, leading to comparisons that don't make sense or that are particularly difficult. Therefore, annotators can also move or resize the mention spans they are annotating.

We use string offsets of mention span pairs to tokens to check that they do not overlap with each other in order to prevent the creation of duplicates. Figure 1 shows an illustrated example of the generation pipeline for mention pairs.

3.5 Annotation Protocol

We recruited three university-educated human annotators and provided them with detailed annotation guidelines for the resolution of yes/no questions on potentially co-referring entities in pairs from the ordered queue described above. By default each entity pair resolution is carried out once, allowing us to quickly expand our data set. However, we pseudo-randomly sample 5% of men-

---
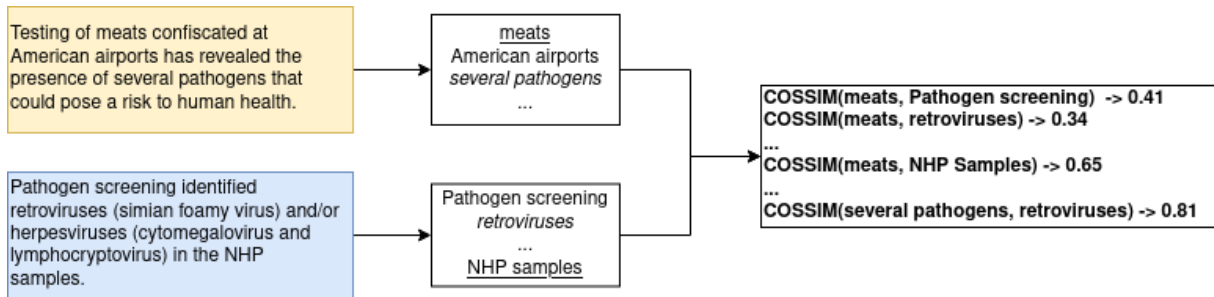
[5]https://github.com/ravenscroftj/cdcrtool

273

Figure 1: Illustration of the generation process for pairs of potentially co-referring expressions, left boxes represent related news summary (top) and abstract (bottom), co-referent entity pairs in middle boxes shown with same formatting (underline,italic).

tion pairs in order to calculate inter-annotator-agreement (IAA) and make sure that data collected from the tool is consistent and suitable for modelling. New entity pairs for IAA are continually sampled as new document pairs and mention tuples are added to the corpus by the web scraper (Section 3.1). The annotation system puts mention pairs flagged for IAA first in the annotation queue. Thus, all annotators are required to complete IAA comparisons before moving on to novel mention pairs. This allows us to ensure that all annotators are well represented in the IAA exercise. To avoid annotators being faced with a huge backlog of IAA comparisons before being able to proceed with novel annotations, we also limited the number of comparisons for IAA required by each user to a maximum of 150 per week.

3.6 Task Difficulty and Annotator Agreement

We anticipated that annotation of the CD^{2}CR corpus would be difficult in nature due to its dependencies on context and lexical style. We invited users to provide feedback regularly to help us refine and clarify our guidelines and annotation tool in an iterative fashion. Users could alert us to examples they found challenging by flagging them as difficult in the tool. Qualitative analysis of the subset of 'difficult' cases showed that the resolution of mention pairs is often perceived by annotators as difficult when:

- Deep subject-matter-expertise is required to understand the mentions, e.g. is "jasmonic acid" the same as "regulator cis -(+)-12-oxophytodienoic acid".

- Mentions involve non-commutable set membership ambiguity e.g. "Diplodocidae" and "the dinosaurs"

- Mentions are context dependent e.g. "the struggling insect" and "the monarch butterfly".

This feedback prompted the introduction of highlighting for existing co-reference chains in the user interface (as described in section 3.4 above) to make it easier to tell when non-commutable set membership would likely introduce inconsistencies into the dataset. For mention pairs requiring subject-matter-expertise, annotators were encouraged to research the terms online. For context sensitive mention pairs, annotators were encouraged to read the full news article and full scientific paper in order to make a decision.

In our 11 page annotation guidelines document (appendix) we describe the use of our annotation tool and illustrate some challenging CD^{2}CR tasks and resolution strategies. For example precise entities mentioned in the scientific document may be referenced using ambiguous exophoric mentions in the news article (e.g. 'a mountain breed of sheep' vs 'eight ovis aries'). Our guidelines require resolving these cases based on the journalist's intent (e.g. 'a mountain breed' refers to the 'ovis aries' sheep involved in the experiment).

We evaluated the final pairwise agreement between annotators using Cohen's Kappa (Cohen, 1960) (κ_{cohen}) and an aggregate 'n-way' agreement score using Fleiss' Kappa (Fleiss, 1971) (κ_{fleiss}). Pairwise κ_{cohen} is shown in table 2 along with the total number of tasks each annotator completed. Annotator 3 (A3) shows the most consistent agreement with the other two annotators. Our Fleiss' Kappa analysis of tasks common across the three annotators gave $\kappa_{\text{fleiss}} = 0.554$. We note that Fleiss' Kappa is a relatively harsh metric and values, like ours, between 0.41 and 0.60 are considered to demonstrate 'moderate agreement'(Landis

274

	# Annotations	A1	A2	A3
A1	10,685	-	0.492	0.600
A2	3,051	0.492	-	0.500
A3	9,847	0.600	0.500	-

Table 2: Number of Annotations and Pairwise Cohen's Kappa scores κ_{cohen} between annotators.

and Koch, 1977). We also carried out Fleiss' Kappa analysis on the subset of mention pairs that were completed by all annotators and were also marked as difficult by at least one user (180 mention pairs in total). We found that for this subset of pairs, $\kappa_{\text{fleiss}} = 0.399$ which is considered to be fair agreement(Landis and Koch, 1977).

4   Model

Below we describe several baseline models including state of the art CDCR models that we used to evaluate how well current approaches can be used in our CD^2CR task setting.

4.1   BERT Cosine Similarity (BCOS) Baseline

In this model we calculate the cosine-similarity between embeddings of the two mentions in context (M_0, M_1) encoded using a pre-trained BERT model as discussed above in section 3.3. We define a thresholding function f to decide if M_0 and M_1 are co-referent ($f(x) = 1$) or not ($f(x) = 0$):

$$f(x) = \begin{cases} 1, & \text{if COSSIM}(M_0, M_1) \geq t \\ 0, & \text{otherwise} \end{cases}$$

During inference, we pass this function over all pairs M_0, M_1 and infer missing links such that if $f(A, B) = 1$ and $f(B, C) = 1$ then $f(A, C) = 1$.

Based on Figure 2, we test values in increments of 0.01 between 0.3 and 0.8 inclusive for threshold cut off t. We evaluated the baseline by measuring its accuracy at predicting co-reference in each mention pair in the CD^2CR development set. The best performance was attained when $t = 0.65$. A visualisation of the BERT Cosine Similarity distributions of co-referent and non co-referent annotated mention pairs can be seen in Figure 2.

Co-referent mention pairs tend to have a slightly higher BERT cosine similarity than non co-referent mention pairs but there is significant overlap of the two distributions suggesting that in many cases BERT similarity is too simplistic a measure.
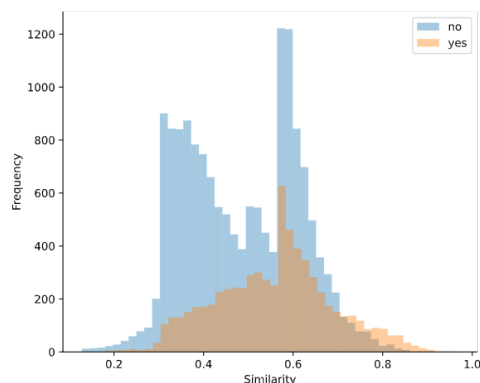

Figure 2: BERT Cosine Similarity frequency distribution for co-referent (Yes) and non-co-referent (No) mention pairs in the CD^2CR corpus.

4.2   Entities Only Baseline (CA)

We use a state-of-the-art model (Cattan et al., 2020) (CA) for cross-document co-reference resolution. In this model, each document is separately encoded using a RoBERTa encoder (without fine-tuning) to get contextualized representations for each token. Then, similarly to the within-document co-reference model by Lee et al. (2017), the mention spans are represented by the concatenation of four vectors: the vectors of the first and last token in the span, an attention-weighted sum of the span token vectors, and a feature vector to encode the span width. Two mention representations are then concatenated and fed to a feed-forward network to learn a likelihood score for whether two mentions co-refer. At inference time, agglomerative clustering is used on the pairwise scores to form coreference clusters.

The CA model is trained to perform both event and entity recognition on the ECB+ corpus (Cybulska and Vossen, 2014) In our setting there is no event detection subtask so, for fair comparison, we pre-train the CA model on ECB+ entity annotations only and evaluate it on our new CD^2CR task to see how well it generalises to our task setting.

4.3   CA + Fine-Tuned (CA-FT) Baseline

Here we aim to evaluate whether fine tuning the CA model from section 4.2 using the CD^2CR corpus can improve its performance in the new task setting. The CA model is first trained on the ECB+ corpus in the manner described above. We then further fine-tune the feed-forward model (without affecting the RoBERTa encoder) on the CD^2CR corpus for 10 epochs with early stopping. Pseudo-random sub-sampling is carried out on the training set to

Are error-correcting code and a two-dimensional array mentions of the same thing?

Yes | No ▾ | Bad Example | This task is hard to think about | Options ▾

Mentions shown highlighted in green are mentions that you have previously annotated as coreferent to one of the two entities.

You can add or remove these secondary mentions using the Options menu.

News Summary[Full Text]

Dr Benjamin Brown from the School of Physics has developed a type of **error-correcting code** for quantum computers that will free up more hardware to do useful calculations. It also provides an approach that will allow companies like Google and IBM to design better quantum microchips. He did this by applying already known code that operates in three-dimensions to a two-dimensional framework.

Science Summary[Full Text]

Fault-tolerant logic gates will consume a large proportion of the resources of a two-dimensional quantum computing architecture. Here we show how to perform a fault-tolerant non-Clifford gate with the surface code; a quantum error-correcting code now under intensive development. This alleviates the need for distillation or higher-dimensional components to complete a universal gate set. The operation uses both local transversal gates and code deformations over a time that scales with the size of the qubit array. An important component of the gate is a just-in-time decoder. These decoding algorithms allow us to draw upon the advantages of three-dimensional models using only **a two-dimensional array** of live qubits. Our gate is completed using parity checks of weight no greater than four. We therefore expect it to be amenable with near-future technology. As the gate circumvents the need for magic-state distillation, it may reduce the resource overhead of surface-code quantum computation considerably.
Task Hash

Figure 3: An example of a cross-document co-reference task presented within our annotation tool.

ensure a balance of co-referent and non-co-referent mention pairs.

4.4 CA - Vanilla (CA-V) Baseline

Here we aim to evaluate whether training the CA model on the CD^{2}CR dataset from the RoBERTa baseline without first training on the ECB+ corpus allows it to fit well to the new task setting. We re-initialise the CA encoder (Section 4.2) using weights from RoBERTa (Liu et al., 2019) and randomly initialise the remaining model parameters. We then train the model on the CD^{2}CR corpus for up to 20 epochs with early stopping with pseudo-random sub-sampling as above.

4.5 CA - SciBERT (CA-S) Baseline

This model is the same as CA-V but we replace the RoBERTa encoder with SciBERT (Beltagy et al., 2019), a version of BERT pre-trained on scientific literature in order to test whether the scientific terms and context captured by SciBERT improve performance at the CD^{2}CR task compared to RoBERTa. Similarly to CA-V in section 4.4, we initialise the BERT model with weights from SciBERT$_{\text{scivocab-uncased}}$ (Beltagy et al., 2019) and randomly initialise the remaining model parameters, training on the CD^{2}CR corpus for up to 20 epochs with early stopping.

5 Results and Discussion

We evaluate each of the model baselines described in section 4 above on the test subset of our CD^{2}CR corpus. Results are shown in table 3.

For the purposes of evaluation, we use named entity spans from the manually annotated CD^{2}CR as

the "gold standard" in all experiments rather than using the end-to-end Named Entity Recognition capabilities provided by some of the models. We evaluate the models using the metrics described by Vilain et al. (1995) (henceforth MUC) and Bagga and Baldwin (1998) (henceforth B^3). MUC F1, precision and recall are defined in terms of pairwise co-reference relationships between each mention. B^3, F1, precision and recall are defined in terms of presence or absence of specific entities in the cluster. When measuring B^3, we remove entities with no co-references (singletons) from the evaluation to avoid inflation of results (Cattan et al., 2020).

The threshold baseline (BCOS) gives the highest MUC recall but also poor MUC precision and poorest B^3 precision. The B^3 metric is highly specific with respect to false-positive entity mentions and strongly penalises BCOS for linking all non-coreferent pairs with $COSSIM(M_0, M_1) \geq 0.65$. Furthermore, Fig. 2 shows that a thresholding strategy is clearly sub-optimal given that there is a significant overlap of co-referent and non-co-referent pairs with only a small minority of pairs at the top and bottom of the distribution that do not overlap.

Model	MUC			B^3		
	P	**R**	**F1**	**P**	**R**	**F1**
BCOS	0.42	**0.94**	0.58	0.01	0.45	0.00
CA	0.41	0.51	0.46	**0.39**	0.33	0.35
CA-V	0.50	0.69	**0.58**	0.35	0.57	**0.44**
CA-FT	0.47	0.71	0.52	0.30	**0.62**	0.41
CA-S	**0.58**	0.46	0.51	0.32	0.53	0.39

Table 3: MUC and B^3 results from running baseline models on CD^{2}CR test subset, BCOS threshold=0.65

Test Type	Co-referent?	Pass Rate & Total Tests	Example test case and outcome for test case
Anaphora and Exophora resolution	Yes	47.1% (16/34)	M1: ...to boost **the struggling insect**'s numbers... *[PASS]* M2: the annual migration of the **monarch butterfly**...
	No	76.5% (26/34)	M1: ...**monarchs** raised in captivity... *[FAIL]* M2: ... wild-caught **monarchs** in an indoor environment...
Subset relationship resolution	Yes	24.3% (9/37)	M1: ...it was in fact a hive of **human activity**... *[FAIL]* M2: ...**Pre-Columbian cultural developments**...
	No	60.0% (18/30)	M1: ... **the carnivore's skull**... *[FAIL]* M2: ... **the gigantic extinct Agriotherium africanum**
Para-phrase resolution	Yes	33.3% (13/39)	M1: ...**a giant short-faced bear**... *[PASS]* M2: ...**the gigantic extinct Agriotherium africanum**...
	No	80.5% (29/36)	M1: ...the energy that **existing techniques** require *[FAIL]* M2: ...the lack of **efficient catalysts** for ammonia synthesis

Table 4: A breakdown of specific tests carried out on CA-V model against three challenging types of relationships found in the CD^2CR corpus. [PASS] or [FAIL] indicates CA-V model correctness. Pass Rate is mathematically equivalent to Recall for test sets.

Therefore, despite its promising MUC F1 score, it is clear that BCOS is not useful in practical terms.

Whilst our thresholding baseline above uses BERT, RoBERTa is used by Cattan et al. (2020) as the basis for their state-of-the-art model and thus for our models based on their work. Although the two models have the same architecture, RoBERTa has been shown to outperform BERT at a range of tasks (Liu et al., 2019). However, as shown in Figure 4, the cosine similarity distribution of mention pair embeddings produced by RoBERTa is compressed to use a smaller area of the potential distribution space compared to that of BERT (Figure 2). This compression of similarities may imply a reduction in RoBERTa's ability to discriminate in our task setting. Liu et al. (2019) explain that their byte-pair-encoding (BPE) mechanism, which expands RoBERTa's sub-word vocabulary and simplifies pre-processing, can reduce model performance for some tasks, although this is not further explored in their work. We leave further exploration of RoBERTa's BPE scheme and its effects on the CD^2CR task setting to future work.

All of the models specifically trained on the CD^2CR corpus (CA-V, CA-FT, CA-S) outperform the CA model by a large margin. Furthermore, the CA-V model (without pre-training on ECB+ corpus) outperforms the CA-FT model (with ECB+ pre-training) by 6% MUC and 3% B^3. These results suggest that the CD^2CR task setting is distinct
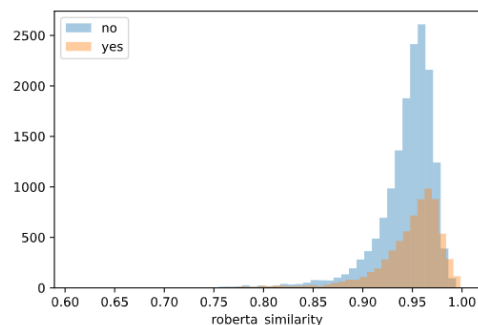

Figure 4: RoBERTa Cosine Similarity frequency distribution for co-referent (Yes) and non-co-referent (No) mention pairs in the CD^2CR corpus. Distribution is compressed between 0.8 and 1.0.

from the CDCR and ECB+ task setting and that this distinction is not solvable with fine-tuning.

In terms of both MUC and B^3, CA-S performs much worse than CA-V suggesting that SciBERT embeddings are less effective than RoBERTa embeddings in this task setting. We hypothesise that SciBERT's specialisation towards scientific embeddings may come at the cost of significantly worse news summary embeddings when compared to those produced by RoBERTa.

We next evaluate our best performing CD^2CR baseline model (CA-V) at the entity resolution CDCR task using the ECB+ test corpus, to see how well it generalises to the original CDCR task. Results are presented in 5 along-side Cattan et al's

277

original model results (CA). The CA-V model still shows good performance, despite a small drop, when compared to the original CA model. The drop in B^3 F1 is more pronounced than MUC but is still broadly in line with other contemporary CDCR systems (Cattan et al., 2020). The CA-V model demonstrates a promising ability to generalise beyond our corpus to other tasks and reveals an interesting correspondence between CDCR and CD^{2}CR settings.

Model	MUC			B^3		
	P	R	F1	P	R	F1
CA	0.86	0.82	0.84	0.63	0.68	0.65
CA-V	0.82	0.81	0.81	0.56	0.53	0.55

Table 5: MUC and B^3 results from running the CD^{2}CR baseline model (CA-V) on ECB+ dataset compared with original Cattan et al. (2020) (CA).

Finally, the best model (CA-V) is analysed using a series of challenging test cases inspired by Ribeiro et al (2020). These test cases were created using 210 manually annotated mention-pairs found in the test subset of the CD^2CR corpus according to the type of relationship illustrated (Anaphora & Exophora, Subset relationships, paraphrases). We collected a balanced set of 30-40 examples of both co-referent and non-coreferent-but-challenging pairs for each type of relationship (exact numbers in Table 4). We then recorded whether the model correctly predicted co-reference for these pairs. The results along with illustrative examples of each relationship type are shown in Table 4. The results suggest that the model is better at identifying non-co-referent pairs than co-referent pairs and that it struggles with positive co-referent mentions for all three types of relationship. The model struggles to relate general reader-friendly descriptions of entities from news articles to precise and clinical descriptions found in scientific papers. The model often successfully identifies related concepts such as 'the carnivore's skull' and 'Agriotherium africanum'. However it is unable to deal with the complexity of these relationships and appears to conflate 'related' with 'co-referent', which is likely due to lack of lexical knowledge as we discussed in section 2.3. Figure 5 shows significant overlap between co-referent and non-co-referent RoBERTa-based cosine similarities, which can also be observed for the wider corpus in Figure 4, but is especially bad for these test examples. This overlap
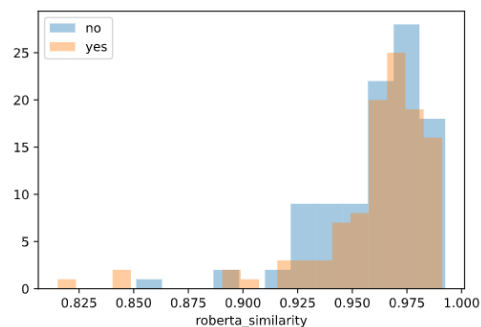

Figure 5: RoBERTa-based mention pair similarity frequency distribution for test examples from Table 4. 'yes' and 'no' for 'co-referent' and 'not-co-referent' respectively

suggests that disentangling these pairs is likely to be a challenging task for the downstream classification layer in the CA-V model. These challenges are less likely to occur in homogeneous corpora like ECB+ where descriptions and relationships remain consistent in detail and complexity.

6   Conclusion

We have defined cross-document, cross-domain co-reference resolution (CD^{2}CR), a special and challenging case of cross-document co-reference resolution for comparing mentions across documents of different types and/or themes. We have constructed a specialised CD^{2}CR annotated dataset, available, along with our annotation guidelines and tool, as a free and open resource for future research. We have shown that state-of-the-art CDCR models do not perform well on the CD^{2}CR dataset without specific training. Furthermore, even with task-specific training, models perform modestly and leave room for further research and improvement. Finally, we show that the understanding of semantic relatedness offered by current generation transformer-based language models may not be precise enough to reliably resolve complex linguistic relationships such as those found in CD^{2}CR as well as other types of co-reference resolution and relationship extraction tasks. The use of semantic enrichment techniques (such as those discussed in Section 2.3) to improve model performance in the CD^{2}CR task should be investigated as future work.

References

Amit Bagga and Breck Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. In *Proceedings of ACL '98/COLING '98*, page 79–85, USA. 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.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. 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 3615–3620.

Arie Cattan, Alon Eirew, Gabriel Stanovsky, Mandar Joshi, and Ido Dagan. 2020. Streamlining cross-document coreference resolution: Evaluation and modeling. arXiv:2009.11032.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

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.

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*, pages 4171–4186. Association for Computational Linguistics.

Sourav Dutta and Gerhard Weikum. 2015. Cross-Document Co-Reference Resolution using Sample-Based Clustering with Knowledge Enrichment. *Transactions of the Association for Computational Linguistics*, 3:15–28.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Matt Grenander, Yue Dong, Jackie Chi Kit Cheung, and Annie Louis. 2019. Countering the Effects of Lead Bias in News Summarization via Multi-Stage Training and Auxiliary Losses. 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 6019–6024.

Zellig S. Harris. 1954. Distributional structure. *WORD*, 10(2-3):146–162.

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

Heng Ji, Ralph Grishman, Hoa Trang Dang, Kira Griffitt, and Joe Ellis. 2010. Overview of the TAC 2010 knowledge base population track. In *Proceedings of the 2010 Text Analysis Conference*.

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.

JR Landis and GG Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159—174.

Anne Lauscher, Ivan Vulić, Edoardo Maria Ponti, Anna Korhonen, and Goran Glavaš. 2020. Specializing unsupervised pretraining models for word-level semantic similarity. *Computing Research Repository*, arXiv:1909.02339.

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.

Belinda Z. Li, Gabriel Stanovsky, and Luke Zettlemoyer. 2020. Active learning for coreference resolution using discrete annotation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8320–8331.

Maria Liakata, Simon Dobnik, Shyamasree Saha, Colin Batchelor, and Dietrich Rebholz-Schuhmann. 2013. A discourse-driven content model for summarising scientific articles evaluated in a complex question answering task. In *Proceedings of Conference on Empirical Methods in Natural Language Processing, EMNLP 2013*, pages 747–757. Association for Computational Linguistics.

Xiao Ling, Sameer Singh, and Daniel S. Weld. 2015. Design challenges for entity linking. *Transactions of the Association for Computational Linguistics*, 3:315–328.

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. *Computing Research Repository*, arXiv:1907.11692.

Annie Louis and Ani Nenkova. 2013. A corpus of science journalism for analyzing writing quality. *Dialogue and Discourse*, 4(2):87–117.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2227–2237.

Edoardo Maria Ponti, Ivan Vulić, Goran Glavaš, Nikola Mrkšić, and Anna Korhonen. 2018. Adversarial propagation and zero-shot cross-lingual transfer of word vector specialization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 282–293.

Jonathan Raiman and Olivier Raiman. 2018. Deep-Type: Multilingual entity linking by neural type system evolution. In *AAAI Conference on Artificial Intelligence*.

Delip Rao, Paul McNamee, and Mark Dredze. 2010. Streaming cross document entity coreference resolution. In *Coling 2010: Posters*, pages 1050–1058, Beijing, China. Coling 2010 Organizing Committee.

James Ravenscroft, Amanda Clare, and Maria Liakata. 2018. HarriGT: A tool for linking news to science. In *Proceedings of ACL 2018, System Demonstrations*, pages 19–24.

Marta Recasens, Lluís Màrquez, Emili Sapena, M. Antònia Martí, Mariona Taulé, Véronique Hoste, Massimo Poesio, and Yannick Versley. 2010. SemEval-2010 task 1: Coreference resolution in multiple languages. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 1–8. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

W. Shen, J. Wang, and J. Han. 2015. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2):443–460.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *MUC6 '95: Proceedings of the 6th conference on Message understanding*, pages 45–52.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. *Computing Research Repository*, arXiv:2004.14974.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. 2013. Ontonotes release 5.0 LDC2013T19. Linguistic Data Consortium, Philadelphia, PA.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. From paraphrase database to compositional paraphrase model and back. *Transactions of the Association for Computational Linguistics*, 3:345–358.

Xiaoyao Yin, Yangchen Huang, Bin Zhou, Aiping Li, Long Lan, and Yan Jia. 2019. Deep entity linking via eliminating semantic ambiguity with BERT. *IEEE Access*, 7:169434–169445.

Mo Yu and Mark Dredze. 2014. Improving lexical embeddings with semantic knowledge. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 545–550.