# Generalizing Cross-Document Event Coreference Resolution Across Multiple Corpora

Michael Bugert
UKP Lab
Department of Computer Science
Technical University of Darmstadt
https://www.ukp.tu-darmstadt.de/

Nils Reimers
UKP Lab

Iryna Gurevych
UKP Lab

*Cross-document event coreference resolution (CDCR) is an NLP task in which mentions of events need to be identified and clustered throughout a collection of documents. CDCR aims to benefit downstream multidocument applications, but despite recent progress on corpora and system development, downstream improvements from applying CDCR have not been shown yet. We make the observation that every CDCR system to date was developed, trained, and tested only on a single respective corpus. This raises strong concerns on their generalizability—a must-have for downstream applications where the magnitude of domains or event mentions is likely to exceed those found in a curated corpus. To investigate this assumption, we define a uniform evaluation setup involving three CDCR corpora: ECB+, the Gun Violence Corpus, and the Football Coreference Corpus (which we reannotate on token level to make our analysis possible). We compare a corpus-independent, feature-based system against a recent neural system developed for ECB+. Although being inferior in absolute numbers, the feature-based system shows more consistent performance across all corpora whereas the neural system is hit-or-miss. Via model introspection, we find that the importance of event actions, event time, and so forth, for resolving coreference in practice varies greatly between the corpora. Additional analysis shows that several systems overfit on the structure of the ECB+ corpus. We conclude with recommendations on how to achieve generally applicable CDCR systems in the future—the most important being that evaluation on multiple CDCR corpora is strongly necessary. To facilitate future research, we release our dataset, annotation guidelines, and system implementation to the public.*[1]

1 https://github.com/UKPLab/cdcr-beyond-corpus-tailored.

Submission received: 25 November 2020; revised version received: 21 February 2021; accepted for publication: 5 June 2021.

https://doi.org/10.1162/COLI_a_00407

## 1. Introduction

To move beyond interpreting documents in isolation in multidocument NLP tasks such as multidocument summarization or question answering, a text understanding technique is needed to connect statements from different documents. A strong contender for this purpose is cross-document event coreference resolution (CDCR). In this task, systems need to (1) find mentions of events in a collection of documents and (2) cluster those mentions together that refer to the same event (see Figure 1). An event refers to an action taking place at a certain time and location with certain participants (Cybulska and Vossen 2014b). CDCR requires deep text understanding and depends on a multitude of other NLP tasks such as semantic role labeling (SRL), temporal inference, and spatial inference, each of which is still being researched and not yet solved. Furthermore, CDCR systems need to correctly predict the coreference relation between any pair of event mentions in a corpus. Because the number of pairs grows quadratically with the number of mentions, achieving *scalable* text understanding becomes an added challenge in CDCR.

In recent years, new CDCR corpora such as the Gun Violence Corpus (GVC) (Vossen et al. 2018) and Football Coreference Corpus (FCC) (Bugert et al. 2020) have been developed, and the state-of-the-art performance on the most commonly used corpus ECB+ (Cybulska and Vossen 2014b) has risen steadily (Kenyon-Dean, Cheung, and Precup 2018; Barhom et al. 2019; Meged et al. 2020). We believe that CDCR can play a vital role for downstream multidocument tasks, and so do other researchers in this area (Bejan and Harabagiu 2014; Yang, Cardie, and Frazier 2015; Upadhyay et al. 2016; Choubey and Huang 2017; Choubey, Raju, and Huang 2018; Choubey and Huang 2018; Kenyon-Dean, Cheung, and Precup 2018; Barhom et al. 2019). Yet, despite the progress made so far, we are not aware of a study that demonstrates that using a recent CDCR system is indeed helpful downstream. We make the key observation that all existing

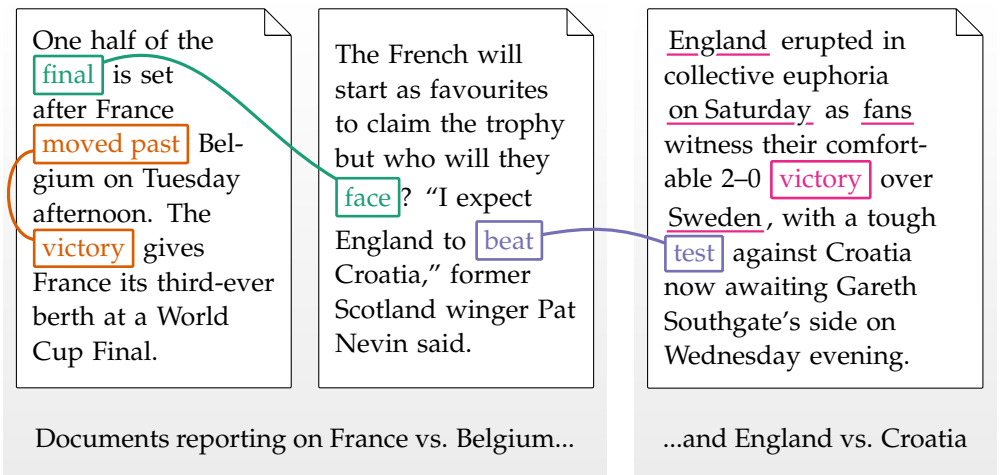


**Figure 1**
Cross-document event coreference resolution (CDCR) example with excerpts of three documents from our token-level reannotation of the Football Coreference Corpus (FCC-T). The seven indicated event mentions refer to four different events. For the *"victory"* event mention, three participant mentions and one temporal mention are additionally marked.

CDCR systems (Kenyon-Dean, Cheung, and Precup 2018; Mirza, Darari, and Mahendra 2018; Vossen 2018; Barhom et al. 2019; Cremisini and Finlayson 2020; Meged et al. 2020) were designed, trained, and evaluated on a single corpus respectively. This points to a risk of systems overspecializing on their target corpus instead of learning to solve the overall task, rendering such systems unsuitable for downstream applications where generality and robustness is required. The fact that CDCR annotation efforts annotated only a subset of all coreference links to save costs (Bugert et al. 2020) further aggravates this situation.

We are, to the best of our knowledge, the first to investigate this risk. In this work, we determine the state of generalizability in CDCR with respect to corpora and systems, identify the current issues, and formulate recommendations on how CDCR systems that are robustly applicable in downstream scenarios can be achieved in the future. We divide our analysis into five successive stages:

1. Cross-dataset modeling of CDCR is made difficult by annotation differences between the ECB+, FCC, and GVC corpora. We establish compatibility by annotating the FCC-T, an extension of the FCC reannotated on the token level.
2. Analyzing generalizability across corpora is best performed with an interpretable CDCR system that is equally applicable on all corpora. To fulfill this requirement, we develop a conceptually simple mention-pair CDCR system that uses the union of features found in related work.
3. To compare the generalization capabilities of CDCR *system architectures*, we train and test this system and a close to state-of-the-art neural system (Barhom et al. 2019) on the ECB+, FCC-T, and GVC corpora. We find that the neural system does not robustly handle CDCR on all corpora because its input features and architecture require ECB+-like corpora.
4. There is a lack of knowledge on how the CDCR task manifests itself in each corpus, especially with regard to which pieces of information (out of event action, participants, time, and location) are the strongest signals for event coreference. Via model introspection, we observe significant differences between corpora, finding that decisions in ECB+ are strongly driven by event actions whereas FCC-T and GVC are more balanced and additionally require text understanding of event participants and time.
5. Finally, we evaluate our feature-based system in a cross-dataset transfer scenario to analyze the generalization capabilities of trained CDCR models. We find that models trained on a single corpus do not perform well on other unseen corpora.

Based on these findings, we conclude with recommendations for the evaluation of CDCR, which will pave the way for more general and comparable systems in the future. Most importantly, the results of our analysis unmistakably show that evaluation on multiple corpora is imperative given the current set of available CDCR corpora.

*Article Structure.* The next section provides background information on the CDCR task, corpora, and systems, followed by related work on feature importance in CDCR (Section 3). Section 4 covers the re-annotation and extension of the FCC corpus. We explain the feature-based CDCR system in Section 5, before moving on to a series of

experiments: We compare this system and the neural system of Barhom et al. (2019) in Section 6. In Section 7 we analyze the signals for event coreference in each corpus. Lastly, we test model generalizability across corpora in Section 8. We discuss the impact of these experiments and offer summarized recommendations on how to achieve general CDCR systems in the future in Sections 9 and 10. We conclude with Section 11.

## 2. Background on CDCR

We explain the CDCR task in greater detail, report on the most influential CDCR datasets, and cover notable coreference resolution systems developed for each corpus.

### 2.1 Task Definition

The CDCR task is studied for several domains including news events in (online) news articles, events pertaining to the treatment of patients in physician's notes (Raghavan et al. 2014; Wright-Bettner et al. 2019), or the identification and grouping of biomedical events in research literature (Van Landeghem et al. 2013). In this work, we restrict ourselves to the most explored variant of CDCR in the news domain.

We follow the task definition and terminology of Cybulska and Vossen (2014b). Here, events consist of four event components—an action, several human or non-human participants, a time, and a location. Each of these components can be mentioned in text, that is, an **action mention** would be the text span referencing the action of an event instance. An example is shown in Figure 1, where the rightmost document references a football match between England and Sweden. The action mention for this event is *"victory,"* alongside three entity mentions *"England"* (the population of England), *"fans"* (English football fans), and *"Sweden"* (the Swedish national football team) who took part in the event. The temporal expression *"on Saturday"* grounds the event mention to a certain time, which in this case depends on the date the news article was published on.

Different definitions have been proposed for the relation of event coreference. Efforts such as ACE (Walker et al. 2006) only permit the annotation of identity between event mentions, whereas Hovy et al. (2013) further distinguish subevent or membership relations. Definitions generally need to find a compromise between complexity and ease of annotation, particularly for the cross-document case (see Wright-Bettner et al. [2019] for a detailed discussion). We follow the (comparatively simple) definition of Cybulska and Vossen (2014b), in which two action mentions corefer if they refer to the same real-world event, meaning their actions and their associated participants, time, and location are semantically equivalent. Relevant examples are shown in Figure 1, where all action mentions of the same color refer to the same event. The two steps a CDCR system needs to perform therefore are (1) the detection of event actions and event components and (2) the disambiguation of event actions to produce a cross-document event clustering. A challenging aspect of CDCR is the fact that finding mentions of all four event components in the same sentence is rare, meaning that information may have to be inferred from the document context or, in some cases, it may not be present in the document at all. The second challenge is efficiently scaling the clustering process to large document collections with thousands of event mentions since every possible pair of event mentions could together form a valid cluster.

### 2.2 System Requirements

The requirements that downstream applications place on systems resolving cross-document event coreference can be diverse. We establish high-level requirements that a system performing CDCR on news text should meet:

- Datasets may consist of many interwoven topics. Systems should perform well on a *broad selection of event types* of different properties (punctual events such as accidents, longer-term events such as natural disasters, pre-planned events such as galas or sports competitions).
- To provide high-quality results, systems should *fully support the definition of event coreference* mentioned previously, meaning they find associations between event mentions at a level human readers would be able to by inferring temporal and spatial clues from the document context and reasoning over event action and participants.
- Datasets may consist of a large number of documents containing many event mentions. We expect CDCR systems to be *scalable* enough to handle 100k event mentions in a reasonable amount of time (less than one day on a single-GPU workstation).

### 2.3 Corpora

The corpus most commonly associated with CDCR is **EventCorefBank+ (ECB+)**. Originally developed as the EventCorefBank (ECB) corpus (Bejan and Harabagiu 2010), it was enriched with entity coreference annotations by Lee et al. (2012) to form the Extended EventCorefBank corpus. This corpus was later extended with 500 additional documents by Cybulska and Vossen (2014b) to create the ECB+ corpus. This most recent version contains 982 news articles on 43 topics. The topics were annotated separately, meaning that there are no coreference links across topics. For each topic (e.g., *"bank explosions"*), there are two main events (*"Bank explosion in Oregon 2008"* and *"Bank explosion in Athens 2012"*) and several news documents that report on either of those two events. The set of documents reporting on the same event is commonly referred to as a **subtopic**. ECB+ is the only corpus of those discussed here that does not provide the publication date for each document. It does, however, contain annotations for all four event components as well as additional cross-document entity coreference annotations for participants, time, and location mentions.

The **Football Coreference Corpus (FCC)** (Bugert et al. 2020) contains 451 sports news articles on football tournaments annotated with cross-document event coreference. The annotation was carried out via crowdsourcing and focused on retrieving **cross-subtopic** event coreference links. Following the nomenclature of Bugert et al. (2020), a within-subtopic coreference link is defined by a pair of coreferring event mentions, which originate from two documents reporting about the same overall event. For example, in ECB+, two different news articles reporting about the same bank explosion in Athens in the year 2012 may both mention the event of the perpetrators fleeing the scene. For a **cross-subtopic** event coreference link, two event mentions from articles on *different* events need to corefer. A sports news article summarizing a quarter-final match of a tournament could for example recommend watching the upcoming semifinal, whereas an article written weeks later about the grand final may refer to the same semifinal in an enumeration of a team's past performances in the tournament.

A concrete example is shown in Figure 1, where the mentions *"beat"* and *"test"* corefer while belonging to different subtopics. Cross-subtopic coreference links are a crucial aspect of CDCR since they connect mentions from documents with low content overlap, forming far-reaching coreference clusters that should prove particularly beneficial for downstream applications (Bugert et al. 2020). In FCC, event mentions are annotated only at the sentence level contrary to ECB+ and GVC which feature token level annotations.

The **Gun Violence Corpus (GVC)** (Vossen et al. 2018) is a collection of 510 news articles covering 241 gun violence incidents. The goal was to create a challenging CDCR corpus with many similar event mentions. Each news article belongs to the same topic (gun violence) and only event mentions related to gun violence were annotated (*"kill," "wounded,"* etc.). Cross-subtopic coreference links were not annotated.

Table 1 presents further insights into these corpora. There, we report the total number of event coreference links in each corpus and categorize them by type. Note that in ECB+ and GVC, nearly all cross-document links are of the within-subtopic kind whereas FCC focused on annotating cross-subtopic links. The stark contrast in the number of coreference links between FCC and ECB+/GVC can be attributed to the facts that (1) the number of coreference links grows quadratically with the number of mentions in a cluster and (2) FCC contains clusters with more than 100 mentions (see Figure 2).

While the annotation design of each of these corpora has had different foci, they share commonalities. The structure of each corpus can be framed as a hierarchy with three levels: there are one or more *topics/event types*, which each contain *subtopics/event*

**Table 1**
Comparison of annotations in several CDCR corpora. Values for FCC refer to our cleaned version of the original corpus. FCC-T is our token-level reannotation.

| | ECB+ | GVC | FCC | FCC-T |
|---|---|---|---|---|
| annotation unit | token | token | sentence | token |
| included annotations | | | | |
| event coreference | ✓ | ✓ | ✓ | ✓ |
| entity coreference | ✓ | · | · | · |
| document publ. date | · | ✓ | ✓ | ✓ |
| particip., time, location | ✓ | · | · | ✓ |
| semantic roles | · | · | · | (✓) |
| topics | 43 | 1 | 1 | 1 |
| subtopics per topic | 2 | 241 | 183 | 183 |
| documents | 982 | 510 | 451 | 451 |
| sentences | 16,314 | 9,782 | 14,940 | 14,940 |
| event mentions | 6,833 | 7,298 | 2,374 | 3,563 |
| event clusters | 2,741 | 1,411 | 218 | 469 |
| singletons | 2,019 | 365 | 50 | 185 |
| entities / event comp. | | | | |
| participant | 12,676 | n/a | n/a | 5,937 |
| time | 2,412 | n/a | n/a | 1,439 |
| location | 2,205 | n/a | n/a | 566 |
| event coreference links | 26,712 | 29,398 | 106,479 | 145,272 |
| within-document | 1,636 | 14,218 | 2,344 | 2,662 |
| within-subtopic | 24,816 | 15,180 | 3,972 | 4,561 |
| cross-subtopic | 260 | 0 | 100,163 | 138,049 |
| cross-topic | 0 | 0 | 0 | 0 |

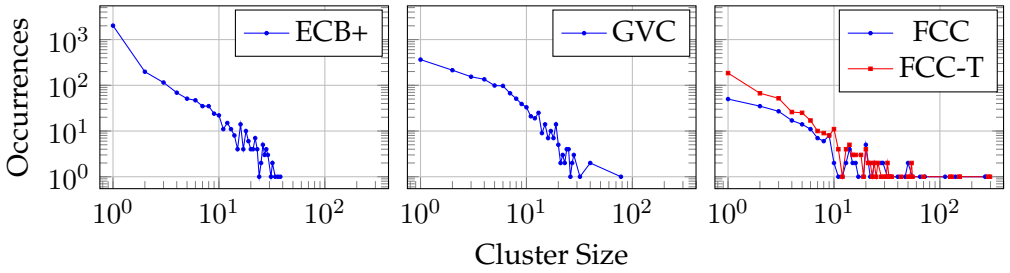


**Figure 2**
Cluster size distribution in CDCR corpora.

*instances*, which each contain multiple *documents*. Both ECB+ and GVC annotate event mentions on the token level in a similar manner. Because FCC is the only CDCR corpus missing token level event mention annotations, we add these annotations in this work to produce the FCC-T corpus (see Section 4). With this change made, it is technically and theoretically possible to examine these CDCR corpora jointly.

### 2.4 Systems

We here summarize the principles of CDCR systems, followed by the state-of-the-art systems for each CDCR corpus.

*2.4.1 System Principles.* Given a collection of event mentions, a discrete or vectorized representation needs to be created for each mention so that the mentions can be clustered. Following the definition of the CDCR task, a representation should contain information on the action, participants, time, and location of the event mention. This information may be scattered throughout the document and needs to be extracted first. To do this, CDCR may preprocess documents via SRL, temporal tagging, or entity linking.

Two general strategies exist for computing the distances between mentions that are needed for clustering: representation learning and metric learning (Hermans, Beyer, and Leibe 2017). **Representation learning** approaches produce a vector representation for each event mention independently. The final event clustering is obtained by computing the cosine distance between each vector pair, followed by agglomerative clustering on the resulting distance matrix. Most approaches belong to the group of conceptually simpler **metric learners**, which predict the semantic distance between two mentions or clusters based on a set of features. By applying the metric on all $\binom{n}{2}$ pairs for $n$ mentions, a distance matrix is obtained that is then fed to a clustering algorithm. Any probabilistic classifier or regression model may be used to obtain the mention distances. Metric learning approaches can be further divided into **mention pair** approaches that compute the distance between each mention pair once and **cluster pair** approaches that recompute cluster representations and distances after each cluster merge. Computing the distance between all mention pairs can be a computationally expensive process. Some metric learner approaches therefore perform a separate document preclustering step to break down the task into manageable parts. The metric learning approach is then applied on each individual cluster of documents and its results are combined to produce the final coreference clustering.

Common types of features used by CDCR systems are text similarity features (string matching between event mention actions), semantic features (the temporal distance

**Table 2**
Preprocessing steps, representations, and features used by CDCR systems. We mark implictly learned neural features with (✓). emb. = embeddings. w.r.t. = with respect to.

| | | CR2020 | ME2020 | BA2019 | KE2018 | VO2016 | CY2015 | YA2015 | LE2012 | MI2018 | VO2018 | **ours** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Prepro-cessing** | Fact KB entity linking | · | · | · | · | ✓ | · | · | · | ✓ | ✓ | ✓ |
| | Lexical KB entity linking | · | · | · | · | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | · |
| | Semantic role labeling | · | ✓ | ✓ | · | ✓ | · | ✓ | ✓ | · | ✓ | ✓ |
| | Temporal tagging | · | · | · | · | ✓ | · | · | · | · | ✓ | ✓ |
| **Mention or Document Represen-tations** | Bag of words | · | · | · | ✓ | · | · | · | ✓ | · | · | · |
| | TF–IDF | ✓ | · | · | ✓ | · | · | ✓ | · | · | ✓ | ✓ |
| | Word emb. simple | ✓ | ✓ | ✓ | ✓ | · | · | ✓ | · | · | · | · |
| | Word emb. contextual | · | ✓ | ✓ | · | · | · | · | · | · | · | ✓ |
| | Character embeddings | · | ✓ | ✓ | · | · | · | · | · | · | · | · |
| **Features** | Entity coreference | · | ✓ | ✓ | · | · | ✓ | · | ✓ | · | · | · |
| | Ling. properties of mention | ✓ | (✓) | (✓) | · | · | · | · | ✓ | · | · | · |
| | Paraphrase detection | · | ✓ | · | · | · | · | · | · | · | · | · |
| | Temporal distance | · | · | · | · | ✓ | · | · | · | ✓ | ✓ | ✓ |
| | Spatial distance | · | · | · | · | · | · | · | · | · | · | ✓ |
| | Compare discrete reprs. | · | · | · | · | ✓ | ✓ | ✓ | ✓ | ✓ | · | ✓ |
| | Compare vectorized reprs. | (✓) | (✓) | (✓) | ✓ | · | · | ✓ | ✓ | · | ✓ | ✓ |
| | Compare w.r.t. lexical KB | · | · | · | · | ✓ | ✓ | · | ✓ | ✓ | · | · |
| | Discourse-related features | · | · | · | ✓ | · | ✓ | · | · | · | · | · |

**Table 3**
Core principles of several CDCR systems.

| System | Target corpus | Mention dist. computation | Approach | Precluster documents? | Learning approach | Clustering technique |
|---|---|---|---|---|---|---|
| CR2020 | ECB+ | classifier | mention pair | yes | MLP | transitive closure |
| ME2020 | ECB+ | classifier | cluster pair | yes | MLP | agglomerative |
| BA2019 | ECB+ | classifier | cluster pair | yes | MLP | agglomerative |
| KE2018 | ECB+ | representation | mention pair | no | MLP autoenc. | agglomerative |
| VO2016 | ECB+ | classifier | mention pair | no | rule-based | transitive closure |
| CY2015 | ECB+ | classifier | mention pair | yes & no | decision tree | transitive closure |
| YA2015 | ECB+ | classifier | mention pair | no | logistic regr. | HDDCRP |
| LE2012 | ECB+ | classifier | cluster pair | yes | linear regr. | agglomerative |
| MI2018 | GVC | classifier | document pair | n/a | rule-based | agglomerative |
| VO2018 | GVC | classifier | document pair | n/a | rule-based | transitive closure |
| **ours** | n/a | classifier | mention pair | no | XGBoost | agglomerative |

between mentions), features using world knowledge (the spatial distance between the locations of mentions), or discourse features (the position of a mention in the document), as well as latent neural features.[2] Table 2 shows the types of features that existing CDCR systems rely on.

*2.4.2 Notable CDCR Systems.* Table 3 shows a comparison of the core principles of several CDCR systems in terms of their mention distance computation, learning approach, and more. We compare the systems of Cremisini and Finlayson 2020 (CR2020), Meged et al. 2020 (ME2020), Barhom et al. 2019 (BA2019), Kenyon-Dean, Cheung, and Precup 2018

2 See Lu and Ng (2018) for more examples of common features.

(KE2018), Vossen and Cybulska 2016 (VO2016), Cybulska and Vossen 2015 (CY2015), Yang, Cardie, and Frazier 2015 (YA2015), Lee et al. 2012 (LE2012), Mirza, Darari, and Mahendra 2018 (MI2018), and Vossen 2018 (VO2018). We emphasize notable systems for each corpus.

At the time of writing, the state-of-the-art system on **ECB+** is Meged et al. (2020), a cluster-pair approach in which a multilayer perceptron is trained to jointly resolve entity and event coreference. It is an extension of Barhom et al. (2019), adding paraphrasing features. The system performs document preclustering prior to the coreference resolution step.

**GVC** was used in SemEval 2018 Task 5 which featured a CDCR subtask (Postma, Ilievski, and Vossen 2018). The best performing system was Mirza, Darari, and Mahendra (2018), which clusters documents using the output of a word sense disambiguation system, person and location entities, and event times. Based on the assumption that each document mentions up to one event of each event type, the system puts all event mentions of the same event type in the same cross-document event coreference cluster. Due to the nature of the shared task, the system is specialized on a limited number of event types. VO2016 and VO2018 are based on the NewsReader pipeline, which contains several preprocessing stages to perform event mention detection, entity linking, word sense disambiguation, and more. Using this information, one rule-based system was defined per corpus (ECB+ and GVC) that is tailored to the topics and annotations present in the respective corpus.

The **FCC** is the most recently released corpus of the three. We are not aware of any publications reporting results for this corpus.

*2.4.3 On the Application of Event Mention Detection.* With respect to the two steps a CDCR system needs to perform (event mention detection and event coreference resolution), several authors have decided to omit the first step and work on gold mentions alone (Cybulska and Vossen 2015; Kenyon-Dean, Cheung, and Precup 2018; Barhom et al. 2019; Meged et al. 2020), which simplifies the task and system development. Systems that include a mention detection step (Lee et al. 2012; Yang, Cardie, and Frazier 2015; Vossen and Cybulska 2016; Choubey and Huang 2017; Vossen 2018; Cremisini and Finlayson 2020) are more faithful to the task but risk introducing another source of error. Compared to using gold event mentions, performance drops from 20 percentage points (pp) CoNLL F1 (Vossen and Cybulska 2016) to 40 pp CoNLL F1 (Cremisini and Finlayson 2020) have been observed on ECB+. Vossen and Cybulska derive from these results that event detection "is the most important factor for improving event coreference" (Vossen and Cybulska 2016, page 518).

We think that the root cause for these losses in performance are not the event detection approaches themselves but rather intentional limitations in the event mention annotations of CDCR corpora. We take the ECB+ corpus as an example. Based on the event definition stated in the annotation guidelines, several hundred event mentions would qualify for annotation in each news document. To keep the annotation effort manageable, only event mentions of the document's seminal event (the main event the article is reporting about) and mentions of other events in the same sentence were annotated (Cybulska and Vossen 2014a, page 9). Conversely, the corpus contains a large amount of valid event mentions that were deliberately left unannotated.[3] A mention detection system will (unaware of this fact) predict these event mentions anyway and

3 In 88% of all sentences in ECB+, no event actions are annotated.

will be penalized for producing false positive predictions. In the subsequent mention clustering step, coreference chains involving these surplus mentions increase the risk of incorrect cluster merges between valid mentions and will overall lead to lower precision. A general purpose mention detection system may perform poorly on the FCC and GVC corpora in similar fashion. For these corpora, affordability of the annotation process was achieved by restricting event mentions to certain action types, which lowers the overall number of to-be-annotated event mentions.

We therefore think that, as long as no CDCR corpus exists in which every single event mention is annotated, event detection and event coreference resolution should be treated separately, meaning that event coreference resolution performance should be reported on gold event mentions. For this reason, and because of the different approaches for limiting the number of event mentions in each of the three corpora, we perform all experiments on gold event mention spans in this work.

## 3. Related Work

Prior work has examined feature importance in CDCR systems. Cybulska and Vossen (2015) tested different combinations of features with a decision tree classifier on ECB+. They find that system performance majorly stems from a lemma overlap feature and that adding discourse, entity coreference, and word sense disambiguation features improves BLANC F1 by only 1 pp. Cremisini and Finlayson (2020) conducted a study in which they built a feature-based mention pair approach for ECB+ to gain deeper insights into the importance of features and on the performance impact of document preclustering. Among four features (fastText [Bojanowski et al. 2017] word embedding similarity between event actions, event action word distribution, sentence similarity, and event action part-of-speech comparison), the embedding similarity feature was found to be the most important by far. The use of document preclustering caused an improvement of 3 pp CoNLL F1, leading Cremisini and Finlayson to encourage future researchers in this field to report experiments with and without document preclustering.

Our work significantly deepens these earlier analyses. Because research on CDCR systems has so far only focused on resolving cross-document event coreference in individual corpora, we tackle the issue of generalizability *across multiple corpora*. We use a *broader* set of features and by comparing two CDCR approaches, while previous work focused on the ECB+ corpus using the aforementioned smaller sets of features. We (1) develop a general feature-based CDCR system, (2) apply it on each of the corpora mentioned above, and (3) analyze the information sources in each corpus that are most informative to cross-document event coreference. We thereby provide the first comparative study of CDCR approaches, paving the way for general CDCR, which will aid downstream multidocument tasks.

## 4. FCC Reannotation

We reannotate the Football Coreference Corpus (FCC) to improve its interoperability with ECB+ and the Gun Violence Corpus (FVC).[4]

The FCC was recently introduced by Bugert et al. (2020) as a CDCR corpus with sentence-level event mention annotations (see Section 2.3). We reannotate all event

4 We additionally conducted an annotation of the missing document publication dates in ECB+, but found that dates could only be manually extracted in half of the corpus documents. We therefore did not include these annotations in our experiments. More details on this annotation effort are reported in Appendix F.

mentions on token level, add annotations of event components, and annotate additional event mentions to produce the **FCC-T** corpus (T for token level). The following sections cover our annotation approach, inter-annotator agreement, and the properties of the resulting corpus.

### 4.1 Annotation Task Definition

In the original FCC annotation, crowd annotators were given a predefined set of events and sentences of news articles to work on. Each sentence had to be marked with the subset of events referenced in the sentence. We take these sentences and annotate the action mention of each referenced event on token level. For each event, we additionally annotate the corresponding participants, time, and location mentions appearing in the same sentence as the action mention. To achieve maximum compatibility with existing corpora, we adopted the ECB+ annotation guidelines (Cybulska and Vossen 2014a).[5] We distinguish between different subtypes of participants (person, organization, etc.), time, and location as done by Cybulska and Vossen (2014a). We do not differentiate between action types because all events (pre-)annotated in FCC should belong to the `OCCURRENCE` type (see Cybulska and Vossen 2014a, page 14). We do not annotate (cross-document) entity coreference. We do annotate a rudimentary kind of semantic roles that we found are crucially missing in ECB+: We instruct annotators to link mentions of participants, time, and location to their corresponding action mention.

While developing the guidelines, we noticed cases where sentence-level mentions are evidently easier to work with than token-level mentions. For example, enumerations or aggregated statements over events (such as *"Switzerland have won six of their seven meetings with Albania, drawing the other."*) are difficult to break down into token-level event mentions. Cases like these are not covered by the ECB+ annotation guidelines and were removed in the conversion process. A similar issue is caused by coordinate structures such as *"Germany beat Algeria and France in the knockout stages,"* where two football match events are referenced by the same verb. To handle these cases, we annotated two separate event mentions sharing the same action mention (*"beat"*). Because superimposed mention spans are not supported by coreference evaluation metrics, we additionally provide a version of the corpus in which these mentions are removed.

In FCC, crowdworkers identified a further 1,100 sentences that mention one or more football-related events outside of the closed set of events they were provided with during the annotation. These event mentions were left unidentified by Bugert et al. (2020). We instructed annotators to manually link each event mention in this extra set of sentences to a database of 40k international football matches[6] and again marked and linked the token spans of actions, participants, times, and locations.

Annotators were given the option to mark sentences they found unclear or that were incorrectly annotated by crowdworkers in the original dataset. We manually resolved the affected sentences on a case-by-case basis.

### 4.2 Annotation Procedure and Results

The annotation was carried out with the INCEpTION annotation tool (Klie et al. 2018). We trained two student annotators on a set of 10 documents. The students were given

5 For details and examples, please refer to the guidelines published at `https://github.com/UKPLab/cdcr-beyond-corpus-tailored`.

6 `https://www.kaggle.com/martj42/international-football-results-from-1872-to-2017/version/5`.

**Table 4**
Inter-annotator agreement ($\alpha_U$).

| | |
|---|---|
| action mentions | 0.80 |
| participants, time, location (spans only) | 0.67 |
| participants, time, location (incl. subtype) | 0.57 |

feedback on their work and afterward annotated a second batch of 22 documents independently. Table 4 shows the inter-annotator agreement on this second batch. We report Krippendorff's $\alpha_U$ (Krippendorff 1995), which measures the agreement in span overlap on character level as the micro average over all documents.

For the annotation of action mention extents, which is the most important step in our re-annotation effort, we reach 0.80 $\alpha_U$, indicating good reliability between annotators (Carletta 1996; Artstein and Poesio 2008). The agreement for the annotation of participants, time, and location is lower, at 0.57 $\alpha_U$. We found that this mostly stems from the annotation of participants: In the guidelines, we specify that annotators should only mark an entity as a participant of an event if it plays a significant role in the event action. The larger and more coarse an event is, the more difficult this decision becomes for annotators. One such case is shown in Example 1, where it is debatable if *"Christian Teinturier"* is or is not significantly involved in the tournament event.

**Example 1**
*"Earlier today, French Football Federation vice-president Christian Teinturier said if there was any basis to the reports about Anelka then he should be sent home from the* ***tournament*** *immediately."*

A second reason is that we do not annotate entity coreference, so only a single entity mention is meant to be annotated for each entity participating in an event. In case the same entity appears twice in a sentence, we instruct annotators to choose the more specific description. If the candidates are identical in surface form, annotators are meant to choose the one closer (in word distance) to the event action. There remains a level of subjectivity in these decisions, leading to disagreement.

Overall, we concluded that the annotation methodology produced annotations of sufficient quality. The remaining 419 documents were divided among both annotators. The corpus re-annotation required 120 working hours from annotators (including training and the burn-in test). We fixed a number of incorrect annotations in the crowdsourced FCC corpus. For example, we removed several mentions of generic events (*"winning a World Cup final is every player's dream"*) that were incorrectly marked as referring to a concrete event.

Table 1 shows the properties of the resulting FCC-T corpus alongside ECB+, GVC, and our cleaned version of the sentence-level FCC corpus. Compared with the original FCC corpus, our token-level reannotation offers 50% more event mentions and twice as many annotated events. With respect to the SRL annotations, we analyzed how frequently event components of each type were attached to action mentions. We found that 95.7% of action mentions have at least one participant attached, 41.6% at least one time mention, and 15.8% at least one location mention. We mentioned earlier that cases exist where two or more action mentions share the same token span. A total of 340 out of all 3,563 annotated event mentions in FCC-T fall into this category. A further 154 event mentions did not have a counterpart in the event database (such as matches from

national football leagues). We jointly assigned these mentions to the coreference cluster `other_event`.

By creating the FCC-T, a reannotation and extension of FCC on token level, we provide the first CDCR corpus featuring a large body of cross-subtopic event coreference links that is compatible with the existing ECB+ and GVC corpora.[7] This greatly expands the possibilities for CDCR research over multiple corpora, as we will demonstrate in Sections 6 to 8.

## 5. Defining a General CDCR System

Recent CDCR approaches such as neural end-to-end systems or cluster-pair approaches were shown to offer great performance (Barhom et al. 2019), yet their black box nature and their complexity makes it difficult to analyze their decisions. In particular, our goal is to identify which aspects of a CDCR corpus are the strongest signals for event coreference that cannot be adequately investigated with recent CDCR systems. We therefore propose a conceptually simpler mention pair CDCR approach that uses a broad set of handcrafted features for resolving event coreference in different environments. We thus focus on developing an interpretable system, whereas reaching state-of-the-art performance is of secondary importance. This section explains the inner workings of the proposed system.

### 5.1 Basic System Definition

We resolve cross-document event coreference by considering pairs of event mentions. At training time, we sample a collection of training mention pairs. For each pair, we extract handcrafted features with which we train a probabilistic binary classifier that learns the coreference relation between a pair (*coreferring* or *not coreferring*). The classifier is followed by an agglomerative clustering step that uses each pair's coreference probability as the distance matrix. At prediction time, all $\binom{n}{2}$ mention pairs are being classified without prior document preclustering. For the reasons outlined in Section 2.4.3, we choose to omit the mention detection step and work with the gold event mentions of each corpus throughout all experiments.

### 5.2 Pair Generation for Training

We explain three issues that arise when sampling training mention pairs and how we address them in our system.

The straightforward technique for sampling training pairs is to sample a fixed number of all possible coreferring and non-coreferring mention pairs. Due to the sparsity of the CDCR relation, the resulting set of pairs would mostly consist of non-coreferring pairs when using this technique, with the majority of coreferring pairs left unused. This issue has been partially addressed in the past with weighted sampling to increase the ratio of coreferring pairs (Lee et al. 2012; Barhom et al. 2019).

We identified a second issue, namely, the underrepresentation of mention pairs from the long tail, which weighted sampling does not address: We previously established that cluster sizes in corpora are imbalanced (see Figure 2). If all $\binom{n}{2}$ coreferring pairs are generated for each cluster, the generated pairs will largely consist of pairs

7 The Football Coreference Corpus (FCC-T) is available at `https://tudatalib.ulb.tu-darmstadt.de/handle/tudatalib/2305`.

from the largest clusters.[8] Manual inspection revealed that the variation in how events are expressed is limited, with large clusters exhibiting many action mentions with (near-)identical surface forms.[9] Consequentially, with common pair generation approaches, there is a high chance of generating many mention pairs that carry little information for the classifier, while mention pairs from clusters in the long tail are unlikely to be included.

Another issue we have not yet seen addressed in related work is the distribution of link types in the body of sampled pairs: In terms of the number of mention pair candidates available for sampling, the cross-topic link candidates strongly outnumber the cross-subtopic link candidates, who in turn strongly outnumber the within-subtopic link candidates (and so on) by nature of combinatorics. This particularly concerns the large body of non-coreferring pairs. An underrepresentation of one of these types during training could cause deficiencies for the affected type at test time, hence care must be taken to achieve a balanced sampling.

We address these three issues as follows: (1) We use the distribution of cluster sizes in the corpus to smoothly transition from generating all $\binom{n}{2}$ coreferring pairs for the smallest clusters to generating $(n-1)\cdot c$ pairs for the largest clusters, where $c \in \mathbb{R}^+$ is a hyperparameter. (2) For each type of coreference link (within-document, within-subtopic, etc.), we sample up to $k$ non-coreferring mention pairs for each coreferring pair previously sampled for this type. Details on the sampling approach are provided in Appendix H.

### 5.3 Features and Preprocessing

Related work has demonstrated a great variety in the representations and features used to resolve cross-document event coreference (see Table 2), yet it remains unclear which features contribute the most to the coreference resolution performance on each of the three corpora. We therefore chose to implement a series of preprocessing steps and feature extractors that cover the majority of features used in previous systems.

*5.3.1 Preprocessing.* We perform lemmatization and temporal expression extraction with CoreNLP (Chang and Manning 2012; Manning et al. 2014), using document publication dates to ground temporal expressions for GVC and FCC-T. We manually converted complex TIMEX expressions into date and time (so that `2020-01-01TEV` becomes `2020-01-01T19:00`). For ECB+ and GVC where participant, time, and location mentions are not linked to the action mention, we applied the SRL system by Shi and Lin (2019) as implemented in AllenNLP (Gardner et al. 2018). We map spans with labels `ARGM-DIR` or `ARGM-LOC` to the location, `ARGM-TM` to the time, and `ARG0` or `ARG1` to the participants of each respective event mention. For all corpora we perform entity linking to DBPedia[10] via DBPedia Spotlight (Mendes et al. 2011).

*5.3.2 Features.* The list of handcrafted mention pair features includes (1) string matching on action mention spans, (2) cosine similarity of TF–IDF vectors for various text regions, (3) the temporal distance between mentions, (4) the spatial distance between

8 For example, sorting all event clusters of the ECB+ training split by size, the largest cluster with 38 mentions would produce more coreferring pairs than the smallest 50% of clusters produce together.

9 This is particularly pronounced for FCC-T, where events of football tournaments are mostly mentioned by *"tournament"* or *"World Cup."*

10 `https://dbpedia.org` – We used the latest release from 1 April 2020.

event actions based on DBPedia, and (5) multiple features comparing neural mention representations. These include representations of action mentions, embeddings of the surrounding sentence, and embeddings of Wikidata entities that we obtained via the DBPedia entity linking step. Details on each feature are reported in Appendix C.

### 5.4 Implementation Details

We implemented the system using Scikit-learn (Pedregosa et al. 2011). To obtain test predictions, we applied the following steps separately for each corpus: We perform feature selection via recursive feature elimination (Guyon et al. 2002) on the respective development split. We use a random forest classifier tasked with classifying mention pairs as "coreferring" / "not coreferring" as an auxiliary task for this stage. We then identified the best classification algorithm to use as the probabilistic mention classifier. We tested logistic regression, a multi-layer perceptron, a probabilistic SVM, and XGBoost (Chen and Guestrin 2016). We tuned the hyperparameters of each classifier via repeated 6-fold cross-validation for 24 hours on the respective training split.[11] Using the best classifier, we optimized the hyperparameters of the agglomerative clustering step (i.e., the linkage method, cluster criterion, and threshold) for another 24 hours on the training split. For each experiment, we train five models with different random seeds to account for non-determinism. At test time we evaluate each of the five models and report the mean of each evaluation metric.[12]

## 6. Generalizability of CDCR Systems

We train and test two CDCR systems and several baselines separately on the ECB+, FCC-T, and GVC corpora to evaluate how flexibly these systems can be applied to different corpora (i.e., whether their overall design is sufficiently general for resolving cross-document event coreference in each corpus). The two systems are our proposed general system (see Section 5) and the system of Barhom et al. (2019) (BA2019). We chose BA2019 because it is the best-performing ECB+ system for which an implementation is available.

In Sections 6.1 and 6.2, we define evaluation metrics and baselines. We then establish the performance of the feature-based system (Section 6.3) on the three corpora, including a detailed link-level error analysis which we can only perform with this system. In Section 6.4, we explain how we apply BA2019 and compare its results to those of the feature-based system, analyzing the impact of document preclustering on the coreference resolution performance in the process.

### 6.1 Evaluation Metrics

Related work on CDCR has so far only scored predictions with the CoNLL F1 (Pradhan et al. 2014) metric (and its constituent parts MUC [Vilain et al. 1995], $CEAF_e$ [Luo 2005], and $B^3$ [Bagga and Baldwin 1998]). We additionally score predictions with the LEA metric (Moosavi and Strube 2016). LEA is a link-based metric which, in contrast to other metrics, takes the size of coreference clusters into account. The metric penalizes incorrect merges between two large clusters more than incorrect merges of mentions

11 Details on the hyperparameter optimization procedure are provided in Appendix G.

12 N.B. This applies to every result originating from this model throughout this work. We will therefore not point this out further.

from two singleton clusters. As we have shown that cluster sizes in CDCR corpora vary considerably (see Figure 2), this is a particularly important property. LEA was furthermore shown to be more discriminative than the established metrics MUC, $CEAF_e$, $B^3$, and CoNLL F1 (Moosavi and Strube 2016).

### 6.2 Baselines

We report the two commonly chosen baselines `lemma` and `lemma-δ` as well as a new `lemma-time` baseline based on temporal information:

1. `lemma`: Action mentions with identical lemmas are placed in the same coreference cluster.
2. `lemma-δ`: Document clusters are created by applying agglomerative clustering with threshold δ on the TF–IDF vectors of all documents, then `lemma` is applied to each document cluster. For hyperparameter δ, we choose the value that produces the best LEA F1 score on the training split.
3. `lemma-time`: A variant of `lemma-delta` based on document-level temporal information. To obtain the time of the main event described by each document, we use the first occurring temporal expression or alternatively the publication date of each document. We create document clusters via agglomerative clustering where the distance between two documents is defined as the difference of their dates in hours. We then apply `lemma` to each document cluster. Here, the threshold δ represents a duration which is optimized as in `lemma-delta`.

### 6.3 Establishing the Feature-based System

We run in-dataset experiments to determine the performance of the feature-based CDCR approach on each individual corpus. Details on the splits used for each corpus are reported in Appendix D. When generating mention pairs for training, we undersample coreferring pairs (see Section 5.2) using hyperparameters $c = 8$ and $k = 8$. In experiments involving FCC-T, we use $c = 2$ and $k = 8$ to compensate for the large clusters in this corpus. Details on the choice of hyperparameters are provided in Appendix H. On all three corpora, the best mention pair classification results were obtained with XGBoost, which led us to use it for all subsequent experiments with this system.

*6.3.1 Mention Clustering Results.* The results are shown in Table 5. For brevity, we only report cross-document performance. It is obtained by applying the evaluation metrics on modified gold and key files in which all documents were merged into a single meta document (Upadhyay et al. 2016).

As was initially shown by Upadhyay et al. (2016), the `lemma-δ` baseline is a strong baseline on the ECB+ corpus. The feature-based system performs on par with this baseline.

For FCC-T, the optimal δ produces a single cluster of all documents, which leads to identical results for the `lemma` and `lemma-δ` baselines. This is a direct consequence of the fact that in this corpus, the majority of event coreference links connect documents from different subtopics. In contrast to ECB+, where preclustering documents by textual content produces document clusters that are near-identical to the gold subtopics

**Table 5**
In-dataset CDCR results of baselines and the feature-based system. The full set of metrics is reported in Appendix A.

| Corpus | System | CoNLL | LEA | | |
|---|---|---|---|---|---|
| | | F1 | P | R | F1 |
| ECB+ | lemma | 61.9 | 42.8 | 43.5 | 43.1 |
| | lemma-δ | 74.4 | **71.5** | 53.7 | **61.3** |
| | feature-based | **74.8** | 67.9 | **55.1** | 60.8 |
| FCC-T | lemma | 42.9 | **38.4** | 19.9 | 26.2 |
| | lemma-δ | 42.9 | **38.4** | 19.9 | 26.2 |
| | lemma-time | 39.8 | 36.8 | 14.2 | 20.5 |
| | feature-based | **54.3** | 30.4 | **60.4** | **39.8** |
| GVC | lemma | 33.8 | 8.8 | 29.7 | 13.6 |
| | lemma-δ | 50.3 | 43.8 | 28.7 | 34.7 |
| | lemma-time | 51.5 | 53.8 | 27.3 | 36.2 |
| | feature-based | **59.4** | **56.5** | **38.2** | **45.6** |

(Barhom et al. 2019; Cremisini and Finlayson 2020), such a strategy is disadvantageous for FCC-T because the majority of coreference links would be irretrievably lost after the document clustering step. The `lemma-time` baseline performs worse on FCC-T than `lemma-δ`, indicating that the document publication date is less important than the document content. The feature-based approach outperforms the baselines on FCC-T, showing higher recall but lower precision, which indicates a tendency to overmerge clusters.

The lemma baselines perform worse on GVC than on ECB+ in absolute numbers, which can be attributed to the fact that Vossen et al. (2018) specifically intended to create a corpus with ambiguous event mentions. Furthermore, the baseline results show that knowing about a document's publication date is worth more than knowing its textual content (at least for this corpus). The feature-based system mostly improves over the baselines in terms of recall.

Another noteworthy aspect in Table 5 are the score differences between CoNLL F1 and LEA F1. In the within-document entity coreference evaluations performed by Moosavi and Strube (2016) alongside the introduction of the LEA metric, the maximum difference observed between CoNLL F1 and LEA F1 were roughly 10 pp. Our experiments exhibit differences of 14 pp for systems and up to 20 pp for baselines due to imbalanced cluster sizes in CDCR corpora.

*6.3.2 Mention Pair Classifier Results.* To evaluate the probabilistic mention pair classifier in isolation for different corpora and coreference link types, we compute binarized recall, precision, and F1 with respect to gold mention pairs.[13] The results are reported in Table 6. The GVC test split does not contain coreferring cross-subtopic event coreference

13 Note that this approach puts higher weight on large clusters, as these produce a greater number of mention pairs. It is nonetheless the only evaluation approach we are aware of that permits analyzing performance *per link type*. Link-based coreference metrics such as MUC (Vilain et al. 1995) cannot be used as a replacement, as these (1) require a full clustering opposed to one score per pair and (2) by design abstract away from individual links in a system's response.

**Table 6**
Mention pair classifier performance of the feature-based system for each cross-document coreference link type. "Coreferring" is used as the positive class. The "Links" column shows the total number of links (coreferring and non-coreferring) per type and corpus based on which P/R/F1 were calculated.

| Link type | ECB+ | | | | FCC-T | | | | GVC | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Links | P | R | F1 | Links | P | R | F1 | Links | P | R | F1 |
| within-document | 11k | 57.6 | 55.9 | 56.7 | 7k | 53.5 | 56.2 | 54.8 | 7k | 69.1 | 30.6 | 42.4 |
| within-subtopic | 83k | 65.0 | 54.5 | 59.3 | 21k | 52.0 | 48.4 | 50.2 | 10k | 70.9 | 28.9 | 41.0 |
| cross-subtopic | 87k | 0.0 | 0.0 | 0.0 | 518k | 53.7 | 36.3 | 43.3 | 435k | n/a | n/a | n/a |

links; therefore these cells are marked with "n/a." Five links of this type are present in the ECB+ test split, of which none were resolved by the in-dataset ECB+ model. For FCC-T and GVC, the performance in resolving within-document, within-subtopic, and cross-subtopic event coreference links decreases gradually from link type to link type. This suggests that the greater the distance covered by an event coreference link is in terms of the topic-subtopic-document hierarchy of a corpus, the more difficult it becomes to resolve it correctly.

*6.3.3 Error Analysis.* To gain a better understanding of the system's limitations, we manually analyzed predictions of the mention pair classifier. We analyzed five false-positive and five false-negative cases for each link type and corpus (roughly 90 mention pairs in total).

We found that textual similarity between action mentions accounts for a large portion of mistakes on the ECB+ and GVC corpora: Unrelated but similar action mentions caused false-positive cases and, vice versa, coreferring but merely synonymous action mentions led to false-negative cases. The FCC-T model did not resolve coreference well between mentions like *"the tournament," "this year's cup,"* and *"2018 World Cup,"* contributing to false-negative cases. Also, the model showed a tendency of merging event mentions prematurely when the action mention and at least one participant matched (see example 1 in Table 7), which would explain the high recall and low precision results seen in Table 5. For all three models, we noticed misclassifications when a sentence contained multiple event mentions (see example 2 in Table 7). In the given example, it is likely that information from the unrelated *"shot in his leg"* mention leaked into the representation of the *"grazed"* event, which contributed to the incorrect classification. For ECB+, we noticed that the lack of document publication date information makes certain decisions considerably harder. For example, the earthquake events seen in example 3 are unrelated and took place four years apart. Although one could come to this conclusion with geographic knowledge alone (the provinces lie on opposite sides of Indonesia), date information would have made this decision easier.

It is reassuring that many of the shortcomings we found would be fixable with a cluster-level coreference resolution approach, (joint) resolution of entity coreference, injection of corpus-specific world knowledge (a football match must take place between exactly two teams, etc.), or with annotation-specific knowledge (e.g., knowledge of Vossen et al.'s [2018] domain model for the annotation of GVC). Our system could be improved by incorporating these aspects, however at the cost of becoming more corpus-specific and less interpretable.

**Table 7**
Notable misclassifications found during manual error analysis.

| Ex. | Mention context A | Mention context B |
|---|---|---|
| 1 | **FCC-T, cross-subtopic link, false positive** | |
| | Zlatko Dalic's men will be back at the Luzhniki Stadium on Sunday to face France, who glanced a 1–0 **victory** over Belgium on Tuesday thanks to a Samuel Umtiti header. | Belgium claims third place with a 2–0 **win** over England |
| 2 | **GVC, cross-subtopic link, false positive** | |
| | A 66-year-old man was shot in his leg and **grazed** in his arm early Monday while sitting on a park bench in Charles Village, police said. | The victims – except an 18-year-old man who refused medical attention after a bullet **grazed** his left leg – were taken to UAMS Medical Center. |
| 3 | **ECB+, cross-subtopic link, false positive** | |
| | Tuesday, July 2, 2013. A strong **earthquake** struck Indonesia's Aceh province on Tuesday, killing at least one person and leaving two others missing. | THOUSANDS of frightened residents flooded makeshift refugee camps in Indonesia's West Papua province today after two powerful **earthquakes** flattened buildings and killed at least one person. |

### 6.4 Comparison to Barhom et al. (2019)

We test an established CDCR system, the former state-of-the-art neural CDCR approach of Barhom et al. (2019) (BA2019), for its generalization capabilities on the three corpora.

*6.4.1 Experiment Setup.* We trained one model of BA2019 for each corpus. BA2019 can resolve event and entity coreference jointly. For the sake of comparability we only use the event coreference component of this system for all experiments since FCC-T and GVC do not contain entity coreference annotations. We replicate the exact data preprocessing steps originally used for ECB+ on FCC-T and GVC. This includes the prediction of semantic roles with the SwiRL SRL system (Surdeanu et al. 2007). The FCC-T corpus mainly consists of cross-subtopic event coreference links (see Section 4.2). The trainable part of the BA2019 system (mention representation and agglomerative mention clustering) is meant to be trained separately on each subtopic of a corpus. This is because at prediction time, the partitioning of documents into subtopics will already be handled by a foregoing and separate document preclustering step. In order not to put BA2019 at a disadvantage for FCC-T, we train it on three large groups of documents that correspond to the three football tournaments present in the FCC-T training split instead of training it on the actual FCC-T subtopics. For this corpus, we also apply undersampling with the same parameters as for the feature-based system.

*6.4.2 Options for Document Preclustering.* Clustering documents by their content before applying mention-level event coreference boosts performance on the ECB+ corpus (Upadhyay et al. 2016; Choubey and Huang 2017; Barhom et al. 2019; Cremisini and Finlayson 2020). As was recommended by Cremisini and Finlayson (2020), we report results for several document preclustering strategies to better distinguish the source of performance gains or losses. We compare (1) no preclustering, (2) the gold document

**Table 8**
Comparison of CDCR performance depending on preclustering strategy. All results reported use gold event mentions. Results marked with • are taken from Table 5.

| Corpus | System | Preclustering | CoNLL | LEA | | |
|---|---|---|---|---|---|---|
| | | | F1 | P | R | F1 |
| ECB+ | ours | none • | 74.8 | 67.9 | 55.1 | 60.8 |
| | | *k*-means | 75.4 | 68.9 | 55.6 | 61.5 |
| | | gold | 76.3 | 71.6 | 56.4 | 63.1 |
| | BA2019 | *k*-means | 79.2 | 67.2 | 68.0 | 67.6 |
| | | gold | 79.1 | 66.7 | 68.5 | 67.6 |
| FCC-T | ours | none/gold • | 54.3 | 30.4 | 60.4 | 39.8 |
| | BA2019 | none/gold | 48.1 | 39.9 | 27.2 | 32.4 |
| | | *k*-means | 34.6 | 36.9 | 8.6 | 13.9 |
| GVC | ours | none • | 59.4 | 56.5 | 38.2 | 45.6 |
| | | *k*-means | 60.8 | 63.9 | 38.6 | 48.1 |
| | | gold | 62.8 | 66.1 | 41.1 | 50.7 |
| | BA2019 | *k*-means | 73.3 | 58.7 | 74.3 | 65.6 |
| | | gold | 79.8 | 69.5 | 80.9 | 74.8 |

clusters, and (3) the *k*-means clustering approach used in Barhom et al. (2019). The gold document clusters are defined via the transitive closure of all event coreference links. For GVC, this gold document clustering is identical to the corpus subtopics. For the FCC-T test split, the gold clustering is a single cluster containing all documents, which is equivalent to not applying document clustering at all. For ECB+, the gold clustering largely corresponds to the corpus subtopics with the exception of some subtopics that are merged due to cross-subtopic event coreference links. In the *k*-means approach, all $n$ input documents are represented by TF–IDF vectors based on which all possible *k*-means clusterings for $k = 2, \ldots, n$ are created. From these clusterings, the one with the highest silhouette score (Rousseeuw 1987) is used.

*6.4.3 Results.* Due to long runtimes of BA2019,[14] results reported from this system stem only from a single execution. We do not report experiments without preclustering on ECB+ and GVC using this system due to scalability issues caused by the greater number of event mentions in these corpora (see Table 1). The results are shown in Table 8.

We comment on the most remarkable results. The BA2019 system architecture performs well on GVC, reaching 65.6 LEA F1. Compared to the ECB+ results, there is a notable score difference between the *k*-means and gold preclustering variants on this corpus. The reason is the same one that led to the `lemma-time` baseline outperforming `lemma-`$\delta$ on this corpus—preclustering documents by textual content is less effective on a corpus with a single topic, and BA2019 does not make use of document publication date annotations. For FCC-T, applying BA2019 out-of-the-box with *k*-means preclustering performs much worse than when the preclustering step is omitted due to the large amount of cross-subtopic links being cut off.

14 Training and optimizing one model on the FCC-T corpus took 10 days on an Nvidia V100 GPU.

When comparing systems against each other, BA2019 performs better than the feature-based approach on ECB+ and GVC. The opposite is the case for FCC-T, where the neural model shows greatly reduced recall in comparison to the feature model. This is surprising to some extent since BA2019 is a more powerful cluster-level approach compared to the mention pair approach. A plausible explanation for the performance drop on FCC-T is the narrower set of features in BA2019. Notably, this system lacks world knowledge on locations and participants and does not explicitly model temporal information, all of which would make intuitive sense to have for a corpus mentioning a variety of football players and matches happening on specific dates. The next section adds evidence to this intuition by analyzing in greater depth the information necessary for resolving event coreference in each corpus.

With respect to the experiments conducted in this section, we have shown that it cannot be taken for granted that CDCR systems are sufficiently general to perform equally well on different corpora. This concerns both the quality of their results (which can fluctuate) as well as more fundamental aspects such as their computational complexity (which may preclude their applicability). In the concrete case of BA2019, this comes down to the choice of the input features and the dependency on document preclustering.

## 7. Identifying the Signals for Event Coreference

According to the CDCR task definition (see Section 2.1), coreference between a pair of event mentions requires a match between each of their components (action, participants, time, location). We analyze to which extent corpora satisfy this definition in practice, namely, whether inference over all event components is indeed required, or whether certain event components suffice as signals for resolving event coreference. We approach this analysis in two ways: (1) We investigate the most important features per corpus at training time via model introspection, and (2) we mask the mentions of certain event components in the test split and measure the impact on test performance. We explain the two approaches and present their results (Sections 7.1 and 7.2), then jointly discuss their outcome in Section 7.3.

### 7.1 Feature Importance

Our main reason for developing a feature-based system was that, compared to neural systems, it allows one to directly analyze which input information a model is making use of.

In our system architecture, the agglomerative clustering step is preceded by a mention pair classifier, which we found worked best with the decision tree boosting framework XGBoost (see Section 6). For decision trees, feature importance metrics can be derived from trained models. In Table 9, we report the top features selected during feature selection for each corpus. Alongside, we report the importance of each feature at training time according to the *gain* metric of XGBoost.[15]

For ECB+, the selected features only cover event actions and context representations. Few features were selected overall. For FCC-T, event action and temporal information received the greatest attention. There is a notable absence of document-level features, which we attribute to the fact that document similarity is not of prime

15 Gain refers to the gain in accuracy from introducing a split in a decision tree using a particular feature. See `https://xgboost.readthedocs.io/en/latest/tutorials/model.html` for more details.

**Table 9**
The top five selected features per corpus with feature importance (see Section 6). See Appendix C for detailed feature descriptions and Appendix E for the full listing.

| Gain | Feature Name | Short description |
|---|---|---|
| **ECB+** | | |
| 0.342 | is-lemma-identical | Lemma identity between actions |
| 0.248 | surface-form-mlipns-distance | MLIPNS distance on action surface forms |
| 0.130 | action-mention | SpanBERT action similarity |
| 0.085 | is-surface-form-identical | Action surface form identity |
| 0.056 | document-similarity | TF–IDF document similarity |
| **FCC-T** | | |
| 0.302 | surface-form-mlipns-distance | MLIPNS distance on action surface forms |
| 0.145 | is-lemma-identical | Lemma identity between actions |
| 0.098 | distance-closest-overall-level-year | Years between closest temporal expressions |
| 0.081 | is-surface-form-identical | Surface form identity between actions |
| 0.053 | distance-sentence-level-year | Years between temp. exprs. in same sentence |
| **GVC** | | |
| 0.255 | surface-form-mlipns-distance | MLIPNS distance on action surface forms |
| 0.126 | distance-doc.-pub.-level-week | Week distance of doc. publication dates |
| 0.125 | distance-doc.-pub.-level-month | Month distance of doc. publication dates |
| 0.075 | document-similarity | TF–IDF document similarity |
| 0.059 | is-lemma-identical | Lemma identity between actions |

importance for resolving cross-subtopic coreference links. For GVC a large number of features was selected, suggesting that diverse information is required for coreference resolution decisions. The most prominent features cover the event action and document-level information.

### 7.2 Masking of Event Components

We want to analyze the impact of each type of event component at test time. To do so, we create variants of the test data where the spans of certain event mentions are masked. We then predict with the models trained in the in-dataset scenario (Section 6) and measure the score delta.

When masking mention spans, we replace each token with a unique dummy token.[16] This is to ensure that the string similarity between two mentions is entirely random. For action components, we replace all gold annotated mention spans. For participant, time, and location components, we replace gold annotations as well as any additional entities identified by DBpedia Spotlight. Masked spans are also removed from semantic role arguments. For the FCC-T and GVC corpora, we additionally mask the document publication date. This masking approach is not without limitations. In FCC-T and GVC, only a subset of all events was annotated. For participants and actions, the three corpora annotate only the head of the phrase. Both these cases may lead to information-bearing tokens leaking into the masked dataset. We nevertheless believe that our approach is an effective approach for analyzing the impact of specific event components at test time.

The results are shown in Table 10. On the ECB+ corpus, masking event actions has the strongest impact on performance. This is to be expected since the majority of features

16 We use fixed-length random tokens from the set `[a-zA-Z]*5` .

**Table 10**
Impact on test performance when masking spans of certain event components. We report the score deltas of the feature-based system with respect to the scores from Table 5.

| Corpus | Masking | Δ CoNLL | Δ LEA | | |
|---|---|---|---|---|---|
| | | F1 | P | R | F1 |
| ECB+ | action | −13.57 | −9.61 | −16.37 | −14.29 |
| | location | −0.25 | −0.57 | −0.26 | −0.39 |
| | participants | −1.41 | −2.25 | −1.40 | −1.76 |
| | time | +0.11 | +0.10 | +0.45 | +0.31 |
| FCC-T | action | −7.90 | −17.53 | +1.34 | −18.48 |
| | location | −1.46 | −7.41 | +1.74 | −6.31 |
| | participants | −9.73 | −17.18 | +13.65 | −17.31 |
| | pub. date | +0.00 | +0.00 | +0.00 | +0.00 |
| | pub. date + time | −14.31 | −21.09 | +24.96 | −22.97 |
| GVC | action | −19.12 | −17.23 | −22.60 | −23.26 |
| | location | −0.50 | −2.48 | −0.03 | −0.85 |
| | participants | +0.18 | +1.03 | +0.39 | +0.61 |
| | pub. date | −6.64 | −15.44 | −4.00 | −8.27 |
| | pub. date + time | −18.28 | −38.63 | −9.43 | −23.53 |

used by the model are action-related features. For FCC-T, masking intensifies the issue of cluster overmerging. Between action, participants, and time, the drop in LEA F1 is comparable. When interpreting the effects of masking on FCC-T, it is important to keep in mind that all events annotated in this corpus are *planned* events whose time, location, and (to some extent) participants are known in advance. This increases the frequency by which these event components are mentioned in text, and on the flipside should cause stronger losses in performance compared to ECB+ and GVC, which contain a smaller proportion of planned events. The fact that scores drop for FCC-T only when the document publication date *and* temporal expressions are removed indicates that the document publication date was not used for grounding temporal expressions in text. In terms of the GVC, action and time stand out as the event components with the highest impact whereas location and participant information barely contribute to the results. We were surprised to see that with regard to temporal information, temporal expressions in text carry more information than the document publication date. Manual inspection revealed that in sentences like *"Two-year-old child shot in the chest in Palm Harbor,"* the entity linker would frequently misclassify *"Two-year-old"* as a temporal expression instead of a person entity. A portion of the performance loss from masking time expressions may therefore arise from masked participants.

### 7.3 Summary on the Signals for Event Coreference

Answering our initial question, whether established CDCR corpora match the CDCR task definition in that they require inference over each event component, we conclude that this is not the case.

Our experiments demonstrate that CDCR decisions in ECB+ are strongly driven by action mentions. GVC, designed to overcome this shortcoming of ECB+ (Vossen et al. 2018), necessitates inference over action mentions, time, and, to some level, participants,

but does not challenge systems on spatial inference. FCC-T is the most balanced of the three corpora based on the facts that at training time, feature selection yielded a broad selection of features and, at test time, performance decreases similarly when action, location, or time information is removed. Overall, neither corpus requires inference over all four event components which define the cross-document event coreference relation.

This indicates that CDCR systems that focus on solving a single corpus model only a subset of the entire CDCR task, which severely limits their downstream use on unseen data, as this data may reflect the task differently from what was observed at training time. The findings further raise the question to which degree it is possible to resolve cross-document event coreference in the three corpora with a *single* model, as this would be the most desirable usage scenario for applying CDCR downstream. We address this question in the following section.

## 8. Generalizability of Trained CDCR Models

All preceding experiments in this work have addressed the ECB+, FCC-T, and GVC corpora in isolation, training a separate model per corpus. In downstream application scenarios, such a differentation is not possible—here, a CDCR system is expected to resolve cross-document event coreference in a robust manner regardless of the selection of topics or underlying structure of a given collection of documents. To gain insights into which performance to expect in such a scenario, we test models of the feature-based system in a *cross-dataset* transfer scenario on unseen corpora.[17]

Furthermore, recent research on the question answering (QA) task has shown that training systems jointly on multiple datasets can improve model robustness and can boost performance (Fisch et al. 2019; Talmor and Berant 2019; Guo et al. 2021). In this work, we have established compatibility between the ECB+, FCC-T, and GVC corpora and have identified the different ways in which each corpus models event coreference. We test whether benefits similar to those observed for question answering are possible for CDCR by training the feature-based system on *multiple* CDCR corpora.

### 8.1 Experiment Setup

We use the same splits for all corpora as in previous experiments. In ECB+, a number of topics cover sports news or news related to gun violence. We refer to these corpus subsets as ECB+$_{\text{sports}}$ and ECB+$_{\text{guns}}$, respectively, and treat them separately in our experiments. Their contents are shown in Table 11. The ECB+ test split remains unchanged.

When combining two corpora, we use the union of features previously selected during feature selection. The increase in training data leads to an increase in mention pairs, which prolongs the training process. We therefore optimize the hyperparameters of the mention pair classifier and the agglomerative clustering step for three days each.

### 8.2 Results

The results of our experiments, evaluated with LEA, are shown in Table 12. As we have shown in preceding sections, the requirements for resolving cross-document event coreference vary between the corpora, which strongly influences the feature selection and model training processes. We hence expected models trained on a single corpus to

17 We did not test BA2019 in this scenario because of the scalability issues reported in Section 6.4.

**Table 11**
Topics used as distinct subsets of ECB+.

| Corpus Subset | Train | Dev |
|---|---|---|
| ECB+$_{\text{sports}}$ | 5, 7, 10, 25 | 29, 31 |
| ECB+$_{\text{guns}}$ | 3, 8, 16, 18, 22 | 33 |

**Table 12**
Performance of the feature-based system when trained on a single respective corpus (top) vs. multiple corpora at once (bottom). No document preclustering was applied. The rightmost set of columns shows P, R, and F1 aggregated over the three corpora. The full set of metrics is reported in Appendix B.

| Train + Dev Split | | | | | Test Split Performance (LEA) | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | ECB+ | | | FCC-T | | | GVC | | | Harmonic Means | | |
| ECB+ | ECB+$_{\text{sports}}$ | ECB+$_{\text{guns}}$ | FCC-T | GVC | P | R | F1 | P | R | F1 | P | R | F1 | P | R | F1 |
| ✓ | · | · | · | · | **67.9** | 55.1 | **60.8** | 41.4 | 8.5 | 14.1 | 32.1 | 25.3 | 28.3 | 42.8 | 17.1 | 24.5 |
| · | · | · | ✓ | · | 16.0 | **57.2** | 24.9 | 30.4 | **60.4** | 39.8 | 3.6 | **75.0** | 6.8 | 8.0 | **63.3** | 14.1 |
| · | · | · | · | ✓ | 48.8 | 50.2 | 49.5 | 42.2 | 17.7 | 24.9 | **56.5** | 38.2 | **45.6** | **48.5** | 29.2 | 36.4 |
| · | ✓ | · | ✓ | · | 52.1 | 54.3 | 53.1 | 42.7 | 40.2 | **41.3** | 16.9 | 26.6 | 20.6 | 29.5 | 37.1 | 32.8 |
| · | · | ✓ | · | ✓ | 62.5 | 55.2 | 58.6 | 41.5 | 19.9 | 26.9 | 45.0 | 35.4 | 39.7 | 48.1 | 31.1 | 37.8 |
| ✓ | · | · | ✓ | · | 63.3 | 60.7 | 58.0 | **50.9** | 19.7 | 28.4 | 29.7 | 26.7 | 28.1 | 43.4 | 28.1 | 34.1 |
| ✓ | · | · | · | ✓ | 55.4 | 54.8 | 55.1 | 38.0 | 25.5 | 30.5 | 27.4 | 32.8 | 29.8 | 37.1 | 34.1 | 35.5 |
| · | · | · | ✓ | ✓ | 37.6 | 47.7 | 42.0 | 34.1 | 48.5 | 40.0 | 45.7 | 36.6 | 40.6 | 38.5 | 43.5 | **40.9** |
| ✓ | · | · | ✓ | ✓ | 24.3 | 53.6 | 33.4 | 26.5 | 58.8 | 36.5 | 32.3 | 38.3 | 35.0 | 27.3 | 48.5 | 34.9 |

perform poorly when evaluated on unseen corpora. This is confirmed by the top rows of Table 12, where significant gaps between in-dataset and cross-dataset performance can be observed.

When looking at the performance on individual corpora, models trained on multiple corpora perform consistently worse than those trained on a single corpus, with few exceptions (mixing FCC-T with ECB+$_{\text{sports}}$ or GVC during training leads to more balanced LEA recall and precision). However, the best overall result in terms of the whole task (i.e., across all corpora) was achieved with joint training: The model trained on FCC-T and GVC scores 40.9 mean LEA F1 over all corpora, whereas the best single-corpus model trained on GVC alone only reached 36.4. In conclusion, training on multiple corpora did not boost performance on individual corpora. Nevertheless, joint training on multiple corpora has emerged as an important strategy for reaching general CDCR systems.

We have only scratched the surface of joint training for CDCR. Further improvements may be achieved with more sophisticated training approaches, for example, by mixing together different amounts of each corpus (potentially aiming for certain distributions of coreference link types), testing its effects on CDCR systems beyond mention pair approaches or performing training data augmentation with data from other NLP tasks.

## 9. Discussion

Despite its importance for downstream applications, the generalizability of CDCR systems over different corpora has not received attention in the past.

Our experiments showed that a system achieving state-of-the-art-level performance on ECB+ does not consistently produce results of the same quality when trained and tested on other CDCR corpora (see Section 6.4). This raises the suspicion that similar systems that were developed for a single corpus lack the capacity of generalizing to unseen corpora. This suspicion is substantiated when looking at the results of our cross-dataset experiments. We showed that training a general, feature-based CDCR system on a single corpus yields good results on the test split of that respective corpus, whereas performance on other corpora falls short of these results (see Section 8).

This is due to the fact that the ECB+, FCC-T, and GVC corpora test systems on different, yet equally important, parts of the overall task of performing CDCR in news text (cf. our requirements posed in Section 2.2). Beyond established knowledge, such as ECB+ testing systems on a greater number of topics while offering low variation in event instances (Vossen et al. 2018), we found that:

- The distribution of coreference links in each corpus varies significantly, with GVC offering roughly the same number of within-document and within-subtopic links, whereas FCC-T offers many cross-subtopic links (see Table 1).

- Structural differences between corpora (such as the number of subtopics or mentions) can pose a problem for established CDCR techniques such as document preclustering and lead to performance drops (see Section 6.4.2), and expose or amplify scalability issues in systems (see Section 6.4.3).

- Between the corpora, the relevance of the four event components (action, participants, time, location) for resolving cross-document event coreference varies strongly. In particular, ECB+ stands out for requiring inference over event actions almost exclusively (see Section 7).

This means that by designing a system against a single corpus, significant aspects of CDCR are disregarded. Doing so introduces a bias toward corpora with specific properties, which severely limits a system's usefulness for downstream applications on data that exhibits different properties. Therefore, when claiming that a system is capable of resolving cross-document event coreference in the general case, it is imperative to report its performance on *multiple* CDCR corpora to certify its robustness to all aspects of CDCR annotated therein.

Related to this finding is the recent trend of ECB+ systems applying document preclustering prior to mention-level event coreference resolution, which deserves special attention. Throughout this work, we have pointed out that by preclustering documents via TF–IDF, one can reproduce the subtopics of a corpus. At test time, this yields an increase in precision for the resolution of within-document and within-subtopic links but has the downside of precluding the resolution of cross-subtopic or cross-topic links. This downside is negligible on ECB+ because in this corpus, within-document and within-subtopic links outnumber cross-subtopic links by a factor of 100 (see Table 1). Many recent ECB+ systems apply preclustering (Lee et al. 2012; Barhom et al. 2019; Cremisini and Finlayson 2020; Meged et al. 2020), yet scores drop sharply when such a system is applied out-of-the-box on a corpus with a different distribution of coreference

links (see Table 8). The performance boost from document preclustering therefore comes at the cost of an overspecialization on the coreference link distribution in ECB+, which we consider to be a form of overfitting. This is again a strong point for the evaluation of CDCR systems on multiple corpora by which this (inadvertent) oversight in existing systems can be revealed.[18]

### 9.1 Evaluation Recommendations

We summarize our findings with respect to the evaluation of CDCR with four recommendations for future research that pave the way for more general, comparable, and reliably evaluated CDCR systems.

**1. CDCR systems should be tested on more than one corpus.** The ECB+, FCC-T, and GVC corpora each are unique with respect to their topic structure, selection of topics, and distribution of event coreference links, all of which can have an impact on the performance of a CDCR system. Furthermore, the importance of action, participants, time, and location information varies in each corpus. For systems seeking to solve the CDCR task in general (i.e., where the application scenario does not necessitate the choice of one domain-specific corpus), this prompts for joint evaluation on multiple CDCR corpora to reveal a system's strengths and weaknesses. Where possible, the performance for each link type (within-document, within-subtopic, etc.) should be reported.

**2. The LEA evaluation metric (Moosavi and Strube 2016) should be used as an additional performance indicator for CDCR.** This metric was previously shown to be more discriminative than previous coreference resolution metrics such as CoNLL F1 and takes size differences between clusters into account. We showed that cluster sizes in CDCR corpora vary significantly and observed score deltas between CoNLL F1 and LEA F1 of up to 20 pp, which motivates its use for CDCR over CoNLL F1.

**3. In addition to a blind prediction, system performance should be reported when using gold document clusters**. Cremisini and Finlayson (2020) request that future system development efforts should report scores with or without document clustering. We agree with this suggestion and refine it further. Given that in the research area of within-document coreference it is commonplace to report separate scores for mention identification and coreference resolution, we think distinguishing scores obtained with and without knowledge on the gold corpus structure would only make sense for CDCR. Researchers must take care to define the gold document clusters based on the transitive closure of event coreference links (see Section 6.4). Using the topics or subtopics of a corpus for this purpose produces incorrect results since cross-subtopic (or cross-topic) coreference links in the corpus are not taken into account.

**4. Event detection performance should be evaluated carefully on CDCR corpora.** From the point of view of a general-purpose event mention detection system, the event mention annotations in the ECB+, FCC-T, and GVC corpora are incomplete by design (see Section 2.4.3). Care must be taken to not unfairly penalize a system that includes a mention detection step as it may detect valid event mentions for which no gold annotation exists. We recommend computing mention detection performance only on

18 An argument could be made at this point that reporting precision and recall separately for each type of coreference link (as we did in Section 6.3.2) was as insightful as evaluating a system on a variety of corpora with different properties. This is not the case since this such analysis is only possible for the subset of CDCR systems that compute a degree of coreference on link-level. For representation learning approaches or for metric learning approaches with cluster-level features, link-level analysis is not possible.

those sentences that contain gold event mention annotations and reporting coreference resolution performance using the gold event mention annotations as a remedy.

## 10. Future Work

Having established that evaluation on multiple of the currently available corpora is necessary for a reliable performance assessment of CDCR systems, we consider the development of approaches that show consistent performance in such a scenario as the next short- to medium-term goal for this task.

A key challenge will be achieving systems that scale to collections of 10k–100k documents without precluding the resolution of cross-subtopic and cross-topic links. A foundation has already been laid by Kenyon-Dean, Cheung, and Precup (2018), who investigated scalable representation learning approaches for CDCR. Since current corpora consist of less than 1k documents, this may require the annotation of additional corpora. To keep the costs of annotating corpora of such magnitude manageable, novel semi-automatic annotation techniques would be required. Furthermore, the concept of cross-topic event coreference links has not been investigated yet due to a lack of annotated data. Once sufficient robustness and/or scalability has been achieved, use cases for downstream applications of CDCR could be investigated.

## 11. Conclusion

The usefulness of cross-document event coreference resolution for downstream multi-document NLP tasks has not been demonstrated yet. To perform well on unseen data in general, NLP systems need to robustly handle variations in the data they are applied on. For CDCR, multiple corpora with varying properties have been annotated, yet each CDCR system to date was developed, trained, and evaluated on only a single one of them. Besides hurting comparability, this currently allows little conclusions to be drawn on their robustness and generalizability, which contributes to the initially stated problem. We addressed this situation in several ways:

We eliminated the remaining hurdles that rendered joint training and evaluation on multiple CDCR corpora difficult by creating FCC-T, a reannotation and extension of the Football Coreference Corpus (FCC) on token level.

To identify the unique properties of each corpus for resolving event coreference in practice, we developed a mention pair CDCR system with a broad set of handcrafted features and applied it on the EventCorefBank+ (ECB+), FCC-T, and Gun Violence Corpus (GVC) corpora. Using this system, we found that only a subset of all components by which events are commonly defined (action, participants, time, location) are required for resolving CDCR in each corpus in practice. In particular, ECB+ only focuses on resolving event actions whereas GVC and FCC-T are more balanced and additionally demand interpretation of event dates and its participants. Link-level analysis of this system revealed that mention distance (with respect to the topic-subtopic-document hierarchy of a corpus) positively correlates with difficulty in resolving cross-document event coreference links.

In the first uniform evaluation scenario involving multiple CDCR systems and corpora, we compared the neural ECB+ system of Barhom et al. (2019) to the feature-based system. First, we found that the neural system performs well on GVC but is outperformed by the conceptually simpler mention pair approach on FCC-T. Second, we deduced from these experiments that systems that are developed for ECB+ and that apply document preclustering overfit to the link distribution in this corpus.

In brief experiments with joint training on multiple corpora, we achieve a combined LEA F1 of 40.9 across all three corpora with the feature-based system—over 4.5 pp better than the same system trained on either corpus in isolation.

We offered four recommendations for future research on CDCR. Most importantly, we advocate evaluation on multiple corpora after having provided conclusive evidence that evaluating on a single corpus is and was insufficient.

All in all, with our annotation effort, corpus analyses, experiments, and open source implementation, we have laid a solid foundation for future research on robust and general CDCR systems. Achieving such systems then constitutes a big step forward toward CDCR becoming an integral part of the multidocument NLP pipeline.

## Appendix A: Full In-Dataset CDCR Results

**Table A.1**
Full cross-document event coreference resolution results for the in-dataset scenario. We report cross-document performance (all documents merged into one meta document before scoring). The scores of the feature-based system are the mean of five independent runs. Scores for BA2019 stem from a single run.

| Corpus | System | Preclustering | MUC | | | $B^3$ | | | CEAFe | | | CoNLL | LEA | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | P | R | F1 | P | R | F1 | P | R | F1 | F1 | P | R | F1 |
| ECB+ | lemma | n/a | 59.7 | 69.7 | 64.3 | 58.1 | 67.9 | 62.6 | 66.3 | 52.8 | 58.8 | 61.9 | 42.8 | 43.5 | 43.1 |
| | lemma-δ | n/a | 81.5 | 69.1 | 74.8 | 88.4 | 67.7 | 76.7 | 66.1 | 78.3 | 71.7 | 74.4 | 71.5 | 53.7 | 61.3 |
| | feature-based | none | 76.1 | 76.0 | 76.1 | 81.2 | 71.8 | 76.2 | 72.1 | 72.2 | 72.2 | 74.8 | 67.9 | 55.1 | 60.8 |
| | | *k*-means | 77.4 | 76.2 | 76.8 | 82.3 | 71.8 | 76.7 | 72.0 | 73.2 | 72.6 | 75.4 | 68.9 | 55.6 | 61.5 |
| | | gold | 79.2 | 76.6 | 77.9 | 84.7 | 72.0 | 77.8 | 71.7 | 74.5 | 73.1 | 76.3 | 71.6 | 56.4 | 63.1 |
| | BA2019 | *k*-means | 80.7 | 83.5 | 82.1 | 78.3 | 81.4 | 79.8 | 77.5 | 74.2 | 75.8 | 79.2 | 67.2 | 68.0 | 67.6 |
| | | gold | 80.3 | 83.8 | 82.0 | 77.5 | 81.8 | 79.6 | 77.8 | 73.7 | 75.7 | 79.1 | 66.7 | 68.5 | 67.6 |
| FCC-T | lemma | n/a | 66.2 | 58.8 | 62.3 | 59.3 | 33.3 | 42.6 | 19.4 | 31.0 | 23.9 | 42.9 | 38.4 | 19.9 | 26.2 |
| | lemma-δ | n/a | 66.2 | 58.8 | 62.3 | 59.3 | 33.3 | 42.6 | 19.4 | 31.0 | 23.9 | 42.9 | 38.4 | 19.9 | 26.2 |
| | lemma-time | n/a | 64.5 | 50.7 | 56.7 | 64.7 | 27.6 | 38.7 | 17.6 | 37.5 | 23.9 | 39.8 | 36.8 | 14.2 | 20.5 |
| | feature-based | none/gold | 78.3 | 82.7 | 80.4 | 38.3 | 70.8 | 49.2 | 40.4 | 28.2 | 33.2 | 54.3 | 30.4 | 60.4 | 39.8 |
| | BA2019 | none/gold | 84.7 | 49.1 | 62.2 | 83.0 | 36.0 | 50.2 | 20.8 | 67.1 | 31.7 | 48.1 | 39.9 | 27.2 | 32.4 |
| | | *k*-means | 78.3 | 34.1 | 47.5 | 88.3 | 17.9 | 29.8 | 16.6 | 66.2 | 26.5 | 34.6 | 36.9 | 8.6 | 13.9 |
| GVC | lemma | n/a | 52.1 | 57.9 | 54.8 | 18.4 | 44.2 | 26.0 | 28.6 | 16.0 | 20.5 | 33.8 | 8.8 | 29.7 | 13.6 |
| | lemma-δ | n/a | 70.2 | 52.8 | 60.2 | 70.2 | 41.6 | 52.2 | 28.9 | 57.1 | 38.4 | 50.3 | 43.8 | 28.7 | 34.7 |
| | lemma-time | n/a | 82.0 | 51.7 | 63.4 | 85.4 | 40.6 | 55.0 | 25.3 | 62.0 | 36.0 | 51.5 | 53.8 | 27.3 | 36.2 |
| | feature-based | none | 78.1 | 66.3 | 71.7 | 73.6 | 49.9 | 59.5 | 38.2 | 60.9 | 47.0 | 59.4 | 56.5 | 38.2 | 45.6 |
| | | *k*-means | 83.6 | 66.2 | 73.9 | 81.3 | 49.5 | 61.5 | 36.4 | 66.1 | 46.9 | 60.8 | 63.9 | 38.6 | 48.1 |
| | | gold | 85.3 | 68.8 | 76.2 | 82.7 | 51.7 | 63.6 | 38.2 | 67.2 | 48.7 | 62.8 | 66.1 | 41.1 | 50.7 |
| | BA2019 | *k*-means | 85.2 | 89.1 | 87.1 | 66.0 | 81.0 | 72.7 | 66.4 | 54.7 | 60.0 | 73.3 | 58.7 | 74.3 | 65.6 |
| | | gold | 88.9 | 92.0 | 90.4 | 75.5 | 86.2 | 80.5 | 73.9 | 63.9 | 68.5 | 79.8 | 69.5 | 80.9 | 74.8 |

## Appendix B: Full Cross-Dataset CDCR Results

**Table B.1**
Full cross-document event coreference resolution results for the cross-dataset scenario. We report cross-document performance (all documents merged into one meta document before scoring). The presented scores are the mean of five independent runs.

| Train + Dev | | | | | Test | Metrics | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | MUC | | | $B^3$ | | | CEAFe | | | CoNLL | LEA | | |
| ECB+ | $ECB+_{sports}$ | $ECB+_{guns}$ | FCC-T | GVC | | P | R | F1 | P | R | F1 | P | R | F1 | F1 | P | R | F1 |
| | | | | | ECB+ | 76.1 | 76.0 | 76.1 | 81.2 | 71.8 | 76.2 | 72.1 | 72.2 | 72.2 | 74.8 | 67.9 | 55.1 | 60.8 |
| ✓ | · | · | · | · | FCC-T | 62.9 | 44.8 | 52.4 | 71.0 | 21.6 | 33.2 | 18.8 | 47.5 | 27.0 | 37.5 | 41.4 | 8.5 | 14.1 |
| | | | | | GVC | 55.4 | 51.7 | 53.5 | 50.3 | 40.1 | 44.6 | 33.3 | 42.0 | 37.2 | 45.1 | 32.1 | 25.3 | 28.3 |
| | | | | | ECB+ | 57.3 | 91.1 | 70.4 | 22.1 | 89.0 | 35.4 | 72.2 | 20.6 | 32.0 | 45.9 | 16.0 | 57.2 | 24.9 |
| · | · | · | ✓ | · | FCC-T | 78.3 | 82.7 | 80.4 | 38.3 | 70.8 | 49.2 | 40.4 | 28.2 | 33.2 | 54.3 | 30.4 | 60.4 | 39.8 |
| | | | | | GVC | 73.5 | 89.0 | 80.5 | 6.4 | 82.9 | 12.0 | 31.0 | 5.3 | 9.1 | 33.8 | 3.6 | 75.0 | 6.8 |
| | | | | | ECB+ | 64.5 | 82.5 | 72.4 | 63.5 | 78.9 | 70.4 | 68.8 | 45.6 | 54.8 | 65.9 | 48.8 | 50.2 | 49.5 |
| · | · | · | · | ✓ | FCC-T | 65.2 | 56.7 | 60.6 | 62.6 | 31.0 | 41.4 | 22.3 | 37.7 | 28.0 | 43.3 | 42.2 | 17.7 | 24.9 |
| | | | | | GVC | 78.1 | 66.3 | 71.7 | 73.6 | 49.9 | 59.5 | 38.2 | 60.9 | 47.0 | 59.4 | 56.5 | 38.2 | 45.6 |
| | | | | | ECB+ | 67.0 | 81.7 | 73.6 | 64.2 | 77.6 | 70.3 | 73.6 | 54.1 | 62.4 | 68.8 | 52.1 | 54.3 | 53.1 |
| · | ✓ | · | ✓ | · | FCC-T | 75.7 | 74.6 | 75.1 | 54.7 | 51.5 | 53.0 | 33.7 | 36.4 | 35.0 | 54.4 | 42.7 | 40.2 | 41.3 |
| | | | | | GVC | 52.8 | 55.8 | 54.3 | 28.7 | 41.4 | 33.8 | 34.6 | 26.9 | 30.3 | 39.5 | 16.9 | 26.6 | 20.6 |
| | | | | | ECB+ | 72.3 | 81.0 | 76.4 | 76.6 | 77.3 | 76.9 | 71.2 | 60.8 | 65.6 | 73.0 | 62.5 | 55.2 | 58.6 |
| · | · | ✓ | · | ✓ | FCC-T | 65.0 | 59.3 | 62.0 | 59.3 | 33.4 | 42.7 | 23.3 | 34.2 | 27.7 | 44.2 | 41.5 | 19.9 | 26.9 |
| | | | | | GVC | 69.5 | 63.6 | 66.4 | 60.9 | 48.1 | 53.7 | 39.3 | 52.4 | 45.0 | 55.0 | 45.0 | 35.4 | 39.7 |
| | | | | | ECB+ | 73.2 | 75.6 | 74.4 | 77.2 | 71.8 | 74.4 | 72.2 | 69.3 | 70.7 | 73.2 | 63.3 | 53.6 | 58.0 |
| ✓ | · | · | ✓ | · | FCC-T | 71.3 | 58.7 | 64.4 | 71.6 | 31.0 | 43.3 | 28.0 | 54.1 | 36.9 | 48.2 | 50.9 | 19.7 | 28.4 |
| | | | | | GVC | 57.0 | 54.4 | 55.6 | 46.5 | 41.2 | 43.7 | 35.1 | 41.3 | 37.9 | 45.7 | 29.7 | 26.7 | 28.1 |
| | | | | | ECB+ | 68.8 | 81.9 | 74.8 | 68.5 | 78.1 | 73.0 | 73.1 | 56.3 | 63.6 | 70.4 | 55.4 | 54.8 | 55.1 |
| ✓ | · | · | · | ✓ | FCC-T | 65.4 | 66.1 | 65.8 | 50.0 | 39.2 | 44.0 | 25.9 | 24.3 | 25.1 | 44.9 | 38.0 | 25.5 | 30.5 |
| | | | | | GVC | 60.4 | 61.8 | 61.1 | 40.0 | 46.4 | 42.9 | 38.3 | 34.8 | 36.5 | 46.8 | 27.4 | 32.8 | 29.8 |
| | | | | | ECB+ | 59.8 | 81.7 | 69.1 | 50.6 | 78.2 | 61.4 | 68.4 | 38.0 | 48.9 | 59.8 | 37.6 | 47.7 | 42.0 |
| · | · | · | ✓ | ✓ | FCC-T | 70.3 | 78.0 | 73.9 | 39.6 | 61.0 | 48.0 | 38.6 | 16.0 | 22.7 | 48.2 | 34.1 | 48.5 | 40.0 |
| | | | | | GVC | 71.1 | 65.1 | 67.9 | 60.7 | 48.8 | 54.1 | 39.5 | 52.5 | 45.1 | 55.7 | 45.7 | 36.6 | 40.6 |
| | | | | | ECB+ | 56.8 | 90.0 | 69.6 | 32.3 | 87.2 | 47.1 | 65.8 | 19.2 | 29.8 | 48.8 | 24.3 | 53.6 | 33.4 |
| ✓ | · | · | ✓ | ✓ | FCC-T | 73.6 | 84.4 | 78.6 | 29.6 | 70.7 | 41.7 | 35.3 | 7.7 | 12.6 | 44.3 | 26.5 | 58.8 | 36.5 |
| | | | | | GVC | 63.9 | 66.8 | 65.3 | 42.6 | 50.8 | 46.4 | 43.0 | 35.4 | 38.8 | 50.2 | 32.3 | 38.3 | 35.0 |

## Appendix C: Mention Pair Features

**Table C.1**
Feature Overview.

| Type | Features | Description |
|---|---|---|
| Action mention string distance | • `is-surface-form-identical`<br>• `is-lemma-identical`<br>• `surface-form-mlipns-distance`<br>• `surface-form-levenshtein-distance` | String distance on two action mentions. We compare surface form and lemma for identity and compute Levenshtein and MLIPNS (Shannaq and Alexandrov 2010) distances for lexical and phonetic distances. |
| TF–IDF | • `document-similarity`<br>• `surrounding-sentence-similarity`<br>• `sentence-context-similarity` | We fit TF–IDF vectors on all given documents. We then compute the cosine similarity between TF-IDF vectors of text regions belonging to two mention pairs. For the regions we use (1) the full document, (2) the sentence surrounding the event mention, and (3) a sentence window of 5 sentences surrounding the mention (i.e., its context). |
| Sentence embedding similarity | • `surrounding-sentence`<br>• `doc-start` | We compute the cosine similarity between sentence representations of a sentence pair originating from a mention pair. We compare the sentences surrounding each event mention and the first sentence of a mention's document. Sentence representations are computed with the Sentence-BERT framework (Reimers and Gurevych 2019), using the pretrained `distilbert-base-nli-stsb-mean-tokens` model. |
| Action mention embedding similarity | • `action-mention` | We compute a contextualized span representation of the action mention of each event mention and compute the cosine similarity of these representations for each mention pair. Span representations are created from the pretrained SpanBERT large model (Joshi et al. 2020) using a window of five sentences surrounding each event mention. |
| Spatial distance | • `distance-document-level-{geo-hierarchy-match/geodesic-distance}`<br>• `distance-srl-level-{geo-hierarchy-match/geodesic-distance}`<br>• `distance-sentence-level-{geo-hierarchy-match/geodesic-distance}`<br>• `distance-closest-preceding-sentence-level-{geo-hierarchy-match/geodesic-distance}`<br>• `distance-closest-overall-level-{geo-hierarchy-match/geodesic-distance}` | We obtain a location from each mention in five ways: (1) document-level, where we pick the first entity-linked location in a document, (2) SRL-level, where we use semantic role labeling (SRL) to find the linked location expression attached to the mention action, (3) sentence-level, where we use the location expression closest to the mention action in the same sentence, (4) closest-sentence-level, where we use the closest preceding location expression from all previous sentences and (5), a combination which applies (2), (3), (4) in order until a location expression is found.<br>For each location pair, we compute distances in two ways: (1) We compute the geodesic distance between the coordinates of both locations. (2) For each location, we follow the `subdivision` and `country` relations in DBpedia upwards (from more specific to less specific locations) to find a match between the two locations. The earlier |

**Table C.1**
Continued.

| Type | Features | Description |
|---|---|---|
| | | A match is found, the smaller the distance is between the two locations. |
| Temporal distance | • `distance-document-publish-level-{year/month/week/day/hour}`<br>• `distance-document-level-{year/month/week/day/hour}`<br>• `distance-srl-level-{year/month/week/day/hour}`<br>• `distance-sentence-level-{year/month/week/day/hour}`<br>• `distance-closest-preceding-sentence-level-{year/month/week/day/hour}`<br>• `distance-closest-overall-level-{year/month/week/day/hour}` | Computes temporal distance between temporal expressions belonging to a mention pair on different date fields (difference of days, difference of hours, ...). Multiple variants for finding temporal expressions exist: (1) the document publication date (where available), (2) document-level, where we pick the first temporal expression in a document, (3) SRL-level, where we use SRL to find the temporal expression attached to the mention action, (4) sentence-level, where we use the temporal expression closest to the mention action in the same sentence, (5) closest-sentence-level, where we use the closest preceding temporal expression from all previous sentences and (6), a combination which applies (3), (4), (5) in order until a temporal expression is found. |
| Wikidata embedding similarity | • `action-mention`<br>• `semantic-role-args-{mean/variance/min/max}`<br>• `surrounding-sentence-{mean/variance/min/max}`<br>• `sentence-context-{mean/variance/min/max}`<br>• `doc-start-{mean/variance/min/max}` | We obtain Wikidata QIDs for each DBpedia entity and map these to pretrained embeddings from PyTorch-BigGraph (Lerer et al. 2019). For each mention in a mention pair, we look up the vectors of (1) the linked action mention (where available), (2) of all event components, (3) of all linked entities in the surrounding sentence, (4) of all linked entities in a 5-sentence window around the mention and (5) of all linked entities in the first three document sentences. Between each of these groups, we compute the pairwise cosine similarity between all vectors and retain the mean, variance, minimum and maximum similarity, respectively. |

## Appendix D: Corpus Splits

**ECB+:** We used the official splits defined in the corpus readme file and filtered the sentences according to the extra CSV file provided with the corpus. In Section 6.4 we only compare systems also using these splits.

**FCC-T:** We used the tournaments 2010, 2012, and 2014 for training, 2016 for development, and 2018 for testing. We remove all mentions belonging to the `other_event` cluster (see Section 4.2).

**GVC:** No official splits are provided for this corpus. We compiled a list of all 241 gun violence incidents present in the corpus and the mapping from incident to document. We shuffled the incidents and partitioned them randomly so that train, dev, and test contain 70/15/15 percent of all documents, respectively. The list of documents in each split is provided alongside our system implementation. We remove all mentions with cluster ID 0: This cluster contains mentions of generic events or unresolved cross-subtopic coreference links.

## Appendix E: Feature Importance

**Table E.1**
Full list of selected features and feature importance for each corpus.

| Corpus | Gain | Feature | Type |
|---|---|---|---|
| ECB+ | 0.342 | is-lemma-identical | action string distance |
| | 0.248 | surface-form-mlipns-distance | action string distance |
| | 0.130 | action-mention | action mention embedding |
| | 0.085 | is-surface-form-identical | action string distance |
| | 0.056 | document-similarity | TF–IDF |
| | 0.052 | context-similarity | TF–IDF |
| | 0.043 | surface-form-levenshtein-distance | action string distance |
| | 0.028 | surrounding-sentence-similarity | TF–IDF |
| | 0.017 | doc-start | sentence embedding |
| FCC-T | 0.302 | surface-form-mlipns-distance | action string distance |
| | 0.145 | is-lemma-identical | action string distance |
| | 0.098 | distance-closest-overall-level-year | temporal distance |
| | 0.081 | is-surface-form-identical | action string distance |
| | 0.053 | distance-sentence-level-year | temporal distance |
| | 0.052 | surface-form-levenshtein-distance | action string distance |
| | 0.052 | action-mention | action mention embedding |
| | 0.041 | surrounding-sentence-similarity | TF–IDF |
| | 0.032 | semantic-role-args-min | Wikidata embedding |
| | 0.025 | distance-closest-overall-level-week | temporal distance |
| | 0.024 | semantic-role-args-variance | Wikidata embedding |
| | 0.023 | surrounding-sentence-min | Wikidata embedding |
| | 0.020 | semantic-role-args-mean | Wikidata embedding |
| | 0.019 | sentence-context-variance | Wikidata embedding |
| | 0.018 | surrounding-sentence-variance | Wikidata embedding |
| | 0.016 | surrounding-sentence-mean | Wikidata embedding |
| GVC | 0.255 | surface-form-mlipns-distance | action string distance |
| | 0.126 | distance-document-publish-level-week | temporal distance |
| | 0.125 | distance-document-publish-level-month | temporal distance |
| | 0.075 | document-similarity | TF–IDF |
| | 0.059 | is-lemma-identical | action string distance |
| | 0.041 | is-surface-form-identical | action string distance |
| | 0.036 | distance-document-level-week | temporal distance |
| | 0.035 | action-mention | action mention embedding |
| | 0.028 | distance-document-level-month | temporal distance |
| | 0.023 | surface-form-levenshtein-distance | action string distance |
| | 0.020 | doc-start-min | Wikidata embedding |
| | 0.013 | context-similarity | TF–IDF |
| | 0.012 | distance-closest-overall-level-week | temporal distance |
| | 0.011 | doc-start | sentence embedding |
| | 0.009 | distance-closest-overall-level-year | temporal distance |
| | 0.009 | distance-closest-pr-sent-level-week | temporal distance |
| | 0.009 | distance-closest-overall-level-month | temporal distance |
| | 0.009 | distance-closest-pr-sent-level-year | temporal distance |
| | 0.008 | distance-closest-pr-sent-level-month | temporal distance |
| | 0.008 | doc-start-variance | Wikidata embedding |
| | 0.007 | distance-sentence-level-week | temporal distance |
| | 0.007 | distance-document-publish-level-day | temporal distance |
| | 0.007 | doc-start-mean | Wikidata embedding |
| | 0.007 | surrounding-sentence-similarity | TF–IDF |
| | 0.007 | doc-start-max | Wikidata embedding |
| | 0.006 | sentence-context-max | Wikidata embedding |
| | 0.006 | sentence-context-min | Wikidata embedding |
| | 0.006 | distance-closest-pr-sent-level-day | temporal distance |
| | 0.006 | distance-document-level-day | temporal distance |

**Table E.1**
Continued.

| Corpus | Gain | Feature | Type |
|---|---|---|---|
| GVC | 0.006 | `closest-pr-sent-lvl-geo-hier-match` | location |
| | 0.006 | `surrounding-sentence-min` | Wikidata embedding |
| | 0.006 | `surrounding-sentence` | sentence embedding |
| | 0.006 | `sentence-context-mean` | Wikidata embedding |
| | 0.005 | `sentence-context-variance` | Wikidata embedding |

## Appendix F: ECB+ Publication Date Annotation

The document publication date is an important piece of information for grounding temporal expressions, particularly in news text. ECB+ is the only CDCR corpus covered in this work for which document publication dates were not annotated. Source URLs of the corpus articles, from which the publication date could have been extracted automatically, are unfortunately only provided for documents in the ECB+ subtopics added at a later point by Cybulska and Vossen (2014b). For the initial EventCorefBank (ECB), no URLs are present. We manually annotated the missing date by inspecting the first few sentences of each document. Table F.1 displays the number of documents for which the publication date could be identified.

**Table F.1**
Share of documents for which we identified and annotated the publication date. ECB+ refers to the set of documents added by Cybulska and Vossen (2014b) for ECB+.

| | ECB | ECB+ |
|---|---|---|
| Date found | 0 | 460 |
| No date found | 480 | 42 |

Unfortunately, no publication dates were found in the ECB half of the corpus. In light of these results, we decided against using these annotations in our experiments since it may have given systems an unfair advantage in deciding whether a document belongs to one of the ECB or ECB+ subtopics. We nevertheless release our annotations in the hopes that they will be useful for future research.

## Appendix G: Hyperparameter Optimization Procedure

We describe our approach for optimizing the hyperparameters of the feature-based system.

We apply repeated $k$-fold cross-validation to obtain reliable results. To define folds, we first partition the documents based on their topic (ECB+) or subtopic (FCC-T, GVC). From these partitioned document sets, folds are created, based on which we generate mention pairs. Compared to the naïve approach of creating folds from all possible mention pairs of a corpus split, this approach guarantees that *each mention* in the respective test fold is unseen, opposed to just the mention *pair* being unseen (with the two constituent mentions likely having been seen at training time), which provides a more faithful testing scenario. By using topics or subtopics for partitioning, we ensure

1: **function** OPTIMIZE(*model*, set of documents $D$, folds $k$, repetitions $r$)
2: $\varrho^*$, F1$^*$ $\leftarrow$ {}, 0 ▷ Best parameter set and best score
3: $P \leftarrow$ CREATE_DOCUMENT_PARTITION($D$)
4: **while** stopping criterion not met **do**
5: $\varrho \leftarrow$ SAMPLE_HYPERPARAMETER_SET()
6: **for** $rep \leftarrow 1$ to $r$ **do**
7: $folds \leftarrow$ K_RANDOM_FOLDS($P$, $k$) ▷ Each fold is a set of set of documents
8: **for** $i \leftarrow 1$ to $k$ **do**
9: $D_{\text{train}} \leftarrow \bigcup\bigcup_{j \neq i} folds_j$
10: $D_{\text{test}} \leftarrow \bigcup folds_i$
11: $pairs_{\text{train}} \leftarrow$ SAMPLE_TRAINING_MENTION_PAIRS($D_{\text{train}}$)
12: $pairs_{\text{test}} \leftarrow$ GENERATE_ALL_TEST_MENTION_PAIRS($D_{\text{test}}$)
13: F1 $\leftarrow$ EVALUATE(TRAIN(*model*, $\varrho$, $pairs_{\text{train}}$), $pairs_{\text{test}}$)
14: $\overline{\text{F1}} \leftarrow$ mean F1 of all runs performed with $\varrho$
15: $\varrho^*$, F1$^*$ $\leftarrow$ $\varrho$, $\overline{\text{F1}}$ if improvement achieved
16: **return** $\varrho^*$

**Figure G.1**
Hyperparameter optimization approach for the mention pair classifier.

a high number of coreferring pairs in the folds and guide the hyperparameter search toward models that should generalize better across topics or subtopics.

The optimization algorithm is shown in Figure G.1. We use the optuna framework (Akiba et al. 2019) for sampling increasingly optimal sets of hyperparameters and use a configurable maximum duration as the stopping criterion. Optimization of the agglomerative clustering step is performed similarly, with the difference of generating a test clustering in line 12 and using the LEA F1 metric instead of F1 for binary classification in line 13.

### Appendix H: Mention Pair Generation at Training Time

*Approach.* We explain our approach for determining the number of coreferring mention pairs to randomly sample for each event during training. Given a set of events $\mathbb{E} = \{e_1, e_2, \ldots\}$ and the function $\mathrm{m} : \mathbb{E} \rightarrow \mathbb{N}$ providing the number of mentions for an event (i.e., the cluster size), we define $\text{pairs}_{\text{coref}} : \mathbb{E} \rightarrow \mathbb{N}$, the number of coreferring pairs to sample for an event, as:

$$\text{pairs}_{\text{coref}}(e) = \left\lceil (\mathrm{m}(e) - 1) \cdot \min\left(\text{undersample}(e), \frac{\mathrm{m}(e)}{2}\right)\right\rceil \tag{H.1}$$

$$\text{undersample}(e) = c + \mathrm{m}(e)^{1-\text{cdf}(\mathrm{m}(e))} - 1 \tag{H.2}$$

where $c \in \mathbb{R}$ is a hyperparameter and $\text{cdf} : \mathbb{N} \rightarrow \mathbb{Q}$ is the percentage of all mentions in $\mathbb{E}$ originating from clusters up to a given size $i$:

$$\text{cdf}(i) = \frac{\sum_{\{e \in \mathbb{E} | \mathrm{m}(e) \leq i\}} \mathrm{m}(e)}{\sum_{e \in \mathbb{E}} \mathrm{m}(e)} \tag{H.3}$$

For the event $e_{\text{largest}}$ with the most mentions in the given data split, this results in $\text{cdf}(\text{m}(e_{\text{largest}})) = 1$, therefore $\text{pairs}_{\text{coref}}(e_{\text{largest}}) = \lceil(\text{m}(e) - 1) \cdot c\rceil$. Hence, $c$ controls the amount of coreferring mention pairs sampled from large clusters. The number of pairs to sample transitions smoothly from linear to quadratic the smaller a cluster is with respect to the overall distribution of cluster sizes in the dataset.

Having sampled coreferring pairs for each event, we determine their coreference link type (within-document, within-subtopic, etc.). For each type of coreference link (within-document, within-subtopic, etc.), the maximum number of non-coreferring pairs generated is $k$ times the number of generated coreferring pairs. We ensure that the number of non-coreferring pairs increases per link type (so that within-document $<$ within-subtopic $< \ldots$).

*Impact of Hyperparameters.* To evaluate how hyperparameters $c$ and $k$ affect performance, we perform exhaustive grid search with different choices of $c$, $k$ when training the mention pair classifier component of our feature-based approach on the ECB+ training split and evaluating it on the ECB+ development split. We test $c \in \{2^{-3}, 2^{-2}, \ldots, 2^5\}$ and $k \in \{1, 2, 4, 8, 12, 16, 24, 32, 48\}$. The entire training split can produce $7.2 \cdot 10^6$ pairs. With the smallest choice of ($c = 2^{-3}, k = 1$), $0.28 \cdot 10^6$ pairs are generated (just $4 \cdot 10^3$ pairs when excluding non-coreferring cross-topic pairs), whereas the largest choice ($c = 2^5, k = 48$) yields $0.47 \cdot 10^6$ pairs ($0.2 \cdot 10^6$ pairs when excluding non-coreferring cross-topic pairs).

For each $(c, k)$ combination, we compute precision, recall, and F1 for each type of coreference link (using the mean of five independent trials to mitigate noise). We further aggregate these results by computing the macro-average over the four link types to produce one precision/recall/F1 value each per $(c, k)$ tuple, shown in Figure H.1.

As visualized by the plots, $k$ controls the precision/recall tradeoff, with higher $k$ (larger proportion of non-coreferring training pairs) leading to high precision but low recall. The choice of $c$ has a smaller impact on performance. Overall, considering F1 scores, the amount of coreferring mention pairs generated from large clusters can be reduced significantly (with $c$ chosen as low as $2^{-3}$) without loss in performance, unless many non-coreferring pairs are used (high $k$). This indicates that, for ECB+, there is little benefit in generating all possible coreferring mention pairs for training, and that achieving a broad selection of mention pairs from many different events is more important. Based on these results, and taking into account the distribution of cluster sizes in each corpus (see Figure 2), we chose ($c = 8, k = 8$) for ECB+ and the similarly distributed GVC in the main experiments. For experiments involving FCC-T, we chose ($c = 2, k = 8$) to reduce the impact of its few large, mostly redundant clusters on training.

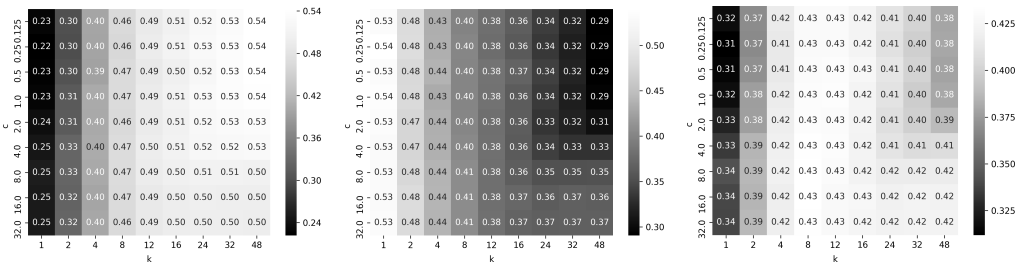


**Figure H.1**
Mention pair classifier performance on the ECB+ development split for different choices of $c$ and $k$. From left to right: Precision, Recall, F1. "Coreferring" is used as the positive class.

## Appendix I: Nomenclature Details

### Acronyms

**CDCR** cross-document event coreference resolution
**SRL** semantic role labeling

### Corpus Acronyms

**ECB** EventCorefBank (The first corpus iteration developed by Bejan and Harabagiu [2010].)
**ECB+** EventCorefBank+ (Extension of the EECB corpus in which Cybulska and Vossen [2014b] added a second subtopic for each topic.)
**EECB** Extended EventCorefBank (Lee et al. [2012]'s extension of the ECB corpus in which entity coreference annotations were added.)
**FCC** Football Coreference Corpus (Sentence-level corpus developed by Bugert et al. [2020].)
**FCC-T** Football Coreference Corpus (Token-level reannotation of FCC produced in this work.)
**GVC** Gun Violence Corpus (Developed by Vossen et al. [2018].)

### System Acronyms

**BA2019** Barhom et al. (2019)
**CR2020** Cremisini and Finlayson (2020)
**CY2015** Cybulska and Vossen (2015)
**KE2018** Kenyon-Dean, Cheung, and Precup (2018)
**LE2012** Lee et al. (2012)
**ME2020** Meged et al. (2020)
**MI2018** Mirza, Darari, and Mahendra (2018)
**VO2016** Vossen and Cybulska (2016)
**VO2018** Vossen (2018)
**YA2015** Yang, Cardie, and Frazier (2015)

### Acknowledgments

We thank Mohsen Mesgar, Kevin Stowe, Prasetya Ajie Utama, and Mingzhu Wu for their helpful comments. Special thanks are due to Jan-Christoph Klie and Nafise Sadat Moosavi for the frequent exchange of ideas. This work was supported by the German Research Foundation through the German–Israeli Project Cooperation (DIP, grant DA 1600/1–1 and grant GU 798/17–1).

### References

Akiba, Takuya, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '19, pages 2623–2631, New York, NY.

Artstein, Ron and Massimo Poesio. 2008. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596. https://doi.org/10.1162/coli.07-034-R2

Bagga, Amit and Breck Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 79–85, Montreal.

Barhom, Shany, 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.

Bejan, Cosmin 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.

Bejan, Cosmin Adrian and Sanda Harabagiu. 2014. Unsupervised event coreference resolution. *Computational Linguistics*, 40(2):311–347. https://doi.org/10.1162/COLI_a_00174

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146. https://doi.org/10.1162/tacl_a_00051

Bugert, Michael, Nils Reimers, Shany Barhom, Ido Dagan, and Iryna Gurevych. 2020. Breaking the subtopic barrier in cross-document event coreference resolution. In *Text2Story@ ECIR*, pages 23–29, Online.

Carletta, Jean. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Chang, Angel X. and Christopher Manning. 2012. SUTime: A library for recognizing and normalizing time expressions. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 3735–3740, Istanbul.

Chen, Tianqi and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16. pages 785–794, ACM, New York, NY, USA.

Choubey, Prafulla Kumar and Ruihong Huang. 2017. Event coreference resolution by iteratively unfolding inter-dependencies among events. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2114–2123, Copenhagen.

Choubey, Prafulla Kumar and Ruihong Huang. 2018. Improving event coreference resolution by modeling correlations between event coreference chains and document topic structures. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 485–495, New Orleans, LA.

Choubey, Prafulla Kumar, Kaushik Raju, and Ruihong Huang. 2018. Identifying the most dominant event in a news article by mining event coreference relations. 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 340–345, New Orleans, LA.

Cremisini, Andres 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.

Cybulska, Agata and Piek Vossen. 2014a. Guidelines for ECB+ annotation of events and their coreference. In *Technical Report*. Technical Report NWR-2014-1, VU University Amsterdam.

Cybulska, Agata and Piek Vossen. 2014b. Using a sledgehammer to crack a nut? Lexical diversity and event coreference resolution. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4545–4552, Reykjavik.

Cybulska, Agata and Piek Vossen. 2015. "Bag of events" approach to event coreference resolution. Supervised classification of event templates. *International Journal of Computational Linguistics and Applications*, 6(2):11–27.

Fisch, Adam, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. 2019. MRQA 2019 shared task: Evaluating generalization in reading comprehension. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 1–13, Hong Kong.

Gardner, Matt, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne.

Guo, Mandy, Yinfei Yang, Daniel Cer, Qinlan Shen, and Noah Constant. 2021. MultiReQA: A cross-domain evaluation for retrieval question answering models. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 94–104, Kyiv.

Guyon, Isabelle, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. 2002. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46:389–422. https://doi.org/10.1023/A:1012487302797

Hermans, Alexander, Lucas Beyer, and Bastian Leibe. 2017. In defense of the triplet loss for person re-identification. *arXiv preprint*.

Hovy, Eduard, Teruko Mitamura, Felisa Verdejo, Jun Araki, and Andrew Philpot. 2013. Events are not simple: Identity, non-identity, and quasi-identity. In *Workshop on Events: Definition, Detection, Coreference, and Representation*, pages 21–28, Atlanta, GA.

Joshi, Mandar, 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. https://doi.org/10.1162/tacl_a_00300

Kenyon-Dean, Kian, 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, LA.

Klie, Jan Christoph, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9.

Krippendorff, Klaus. 1995. On the reliability of unitizing continuous data. *Sociological Methodology*, 25:47–76.

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

Lerer, Adam, Ledell Wu, Jiajun Shen, Timothee Lacroix, Luca Wehrstedt, Abhijit Bose, and Alex Peysakhovich. 2019. PyTorch-BigGraph: A large-scale graph embedding system. In *Proceedings of the 2nd SysML Conference*, Palo Alto, CA.

Lu, Jing and Vincent Ng. 2018. Event coreference resolution: A survey of two decades of research. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 5479–5486, Stockholm.

Luo, Xiaoqiang. 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.

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60, Baltimore, MD.

Meged, Yehudit, Avi Caciularu, Vered Shwartz, and Ido Dagan. 2020. Paraphrasing vs coreferring: Two sides of the same coin. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4897–4907, Online.

Mendes, Pablo N., Max Jakob, Andrés García-Silva, and Christian Bizer. 2011. DBpedia spotlight: Shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems*, I-Semantics '11, pages 1–8, New York, NY.

Mirza, Paramita, Fariz Darari, and Rahmad Mahendra. 2018. KOI at SemEval-2018 task 5: Building knowledge graph of incidents. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 81–87, New Orleans, LA.

Moosavi, Nafise Sadat 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.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Postma, Marten, Filip Ilievski, and Piek Vossen. 2018. SemEval-2018 task 5: Counting events and participants in the long tail. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 70–80, New Orleans, LA.

Pradhan, Sameer, 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, MD.

Raghavan, Preethi, Eric Fosler-Lussier, Noémie Elhadad, and Albert M. Lai. 2014. Cross-narrative temporal ordering of medical events. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 998–1008, Baltimore, MD.

Reimers, Nils 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.

Rousseeuw, Peter J. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Shannaq, Boumedyen and Victor V. Alexandrov. 2010. Using product similarity for adding business value and returning customers. *Global Journal of Computer Science and Technology*, 10:2–8.

Shi, Peng and Jimmy Lin. 2019. Simple BERT models for relation extraction and semantic role labeling. *arXiv preprint, arXiv:1904.05255*.

Surdeanu, Mihai, Lluís Màrquez, Xavier Carreras, and Pere R Comas. 2007. Combination strategies for semantic role labeling. *Journal of Artificial Intelligence Research*, 29:105–151. https://doi.org/10.1613/jair.2088

Talmor, Alon and Jonathan Berant. 2019. MultiQA: An empirical investigation of generalization and transfer in reading comprehension. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4911–4921, Florence.

Upadhyay, Shyam, 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.

Van Landeghem, Sofie, Jari Bjrne, Chih-Hsuan Wei, Kai Hakala, Sampo Pyysalo, Sophia Ananiadou, Hung-Yu Kao, Zhiyong Lu, Tapio Salakoski, Yves Van de Peer, and Filip Ginter. 2013. Large-scale event extraction from literature with multi-level gene normalization. *PLOS ONE*, 8(4):1–12. https://doi.org/10.1371/journal.pone.0055814

Vilain, Marc, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Sixth Message Understanding Conference (MUC-6): Proceedings*, pages 45–52, Columbia, MD.

Vossen, Piek. 2018. NewsReader at SemEval-2018 task 5: Counting events by reasoning over event-centric-knowledge-graphs. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 660–666, New Orleans, LA.

Vossen, Piek and Agata Cybulska. 2016. Identity and Granularity of Events in Text. In *Computational Linguistics and Intelligent Text Processing*, Cham: Springer International Publishing, pages 501–522.

Vossen, Piek, Filip Ilievski, Marten Postma, and Roxane Segers. 2018. Don't annotate, but validate: A data-to-text method for capturing event data. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3034–3042, Paris.

Walker, Christopher, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. ACE 2005 multilingual training corpus. https://catalog.ldc.upenn.edu/LDC2006T06

Wright-Bettner, Kristin, Martha Palmer, Guergana Savova, Piet de Groen, and Timothy Miller. 2019. Cross-document coreference: An approach to capturing coreference without context. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, pages 1–10, Hong Kong.

Yang, Bishan, Claire Cardie, and Peter Frazier. 2015. A hierarchical distance-dependent Bayesian model for event coreference resolution. *Transactions of the Association for Computational Linguistics*, 3:517–528. https://doi.org/10.1162/tacl_a_00155