Breaking the Subtopic Barrier in Cross-Document Event Coreference Resolution

Michael Bugert[♠], Nils Reimers[♠], Shany Barhom[♢], Ido Dagan[♢] and Iryna Gurevych[♠]

[♠] Ubiquitous Knowledge Processing Lab (UKP Lab), Department of Computer Science,
Technical University of Darmstadt, Germany
https://ukp.tu-darmstadt.de
[♢] Bar-Ilan University, Computer Science Department, Ramat Gan, Israel
https://cs.biu.ac.il

Abstract

Cross-document event coreference resolution (CDCR) is the task of detecting and clustering mentions of events across a set of documents. A major bottleneck in CDCR is a lack of appropriate datasets, which stems from the difficulty of annotating data for this task. We present the first scalable approach for annotating *cross-subtopic event coreference links*, a highly valuable but rarely occurring type of cross-document link. The annotation of these links requires combing through hundreds of documents – an endeavor for which conventional token-level annotation schemes with trained expert annotators are too expensive. We instead propose crowdsourcing annotation on sentence level to achieve scalability. We apply our approach to create the Football Coreference Corpus (FCC), a corpus of 451 sports news reports, while reaching high agreement between NLP experts and crowd annotators in the process.[1]

1 Introduction

Events, i.e. actions of participants happening at a specific time and place [CV14], lie at the core of news reporting. Event mention detection is the task of finding spans in text which mention such events. The goal of event coreference resolution is to cluster these event mentions so that each cluster contains mentions referring to the same event. Cross-document event coreference resolution (CDCR) is an extension to multiple documents, producing coreference links within and across document boundaries. Knowing which text passages corefer boosts performance in multi-document downstream tasks such as question answering [Mor99; PIV18] and enables applications such as news timeline generation [Min+15].

CDCR datasets generally consist of a set of pre-clustered documents with added coreference annotations. We follow the terminology of Cybulska and Vossen [CV14] and define a *subtopic* as a cluster of documents reporting about the same event, for example "France beats Croatia to win the FIFA World Cup 2018". To increase diversity, datasets may provide several subtopics from the same domain, referred to as a *topic*. This way, a second set of documents reporting about the event "Croatia beats England in the semifinal of the FIFA World Cup 2018" would form another subtopic. Together with the previous example, the two subtopics are part of the topic "football matches". The event reported in a document which determines its (sub)topic is referred to as the *seminal event* [BH14].

In: R. Campos, A. Jorge, A. Jatowt, S. Bhatia (eds.): Proceedings of the Text2Story'20 Workshop, Lisbon, Portugal, 14-April-2020, published at http://ceur-ws.org

[1] The dataset is available at https://tudatalib.ulb.tu-darmstadt.de/handle/tudatalib/2305.

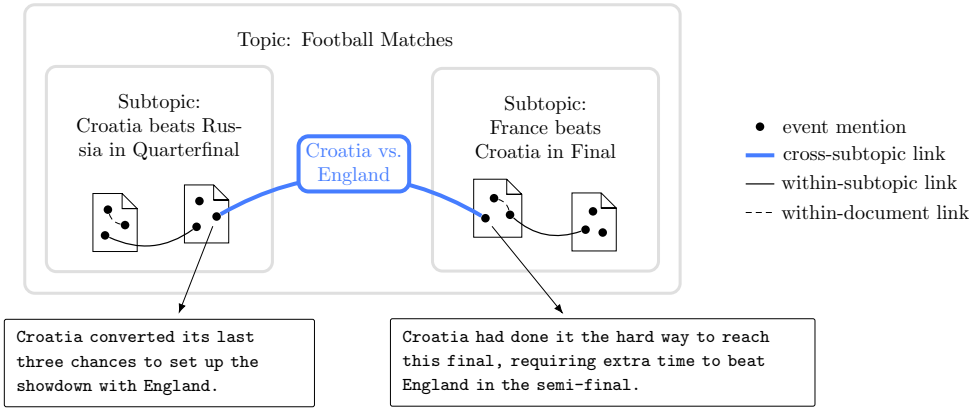


Figure 1: Example of a cross-subtopic event coreference link.

A core technique in journalistic writing is incorporating references to past or future events to establish the context of an article. For example, an article reporting about the World Cup final between France and Croatia may reference the semifinal match Croatia has won in order to reach the final. In the topic and subtopic terminology, such a reference would trigger *cross-subtopic event coreference* – a reference to an event other than the article's seminal event. If one finds a second mention of this event (most likely in a different document), a cross-subtopic event coreference link is established (see Figure 1).

Cross-subtopic links connect documents with partial overlap in action, participants, time or location. Thereby, these links span a large network across documents, making them a particularly vital resource for downstream applications relying on extracting knowledge from multiple documents. Mentions which trigger cross-subtopic event coreference can be found in a handful of locations throughout a given news article. However, due to the limitless set of events that news articles can refer to, finding an established cross-subtopic event coreference *link* is hard and requires sifting through a large set of documents.

The costs for capturing such a sparse phenomenon via conventional token-level event coreference annotation at a large scale would be prohibitive. In particular, ECB+ [CV14], the most frequently used CDCR dataset, contains around 1% of cross-subtopic event coreference clusters, however requiring complex annotation guidelines and trained expert annotators. The same limitations apply to the improved schema used for the creation of the Gun Violence Corpus (GVC) [Vos+18] which offers several hundred subtopics on the topic of gun violence, yet none of its coreference clusters cross the subtopic boundary. To the best of our knowledge, existing annotation schemes are unfit for the task of annotating cross-document event coreference at scale. Furthermore, we are not aware of a CDCR dataset specifically targeting cross-subtopic event coreference links.

To fill this gap, we contribute the following:

1. We propose the first annotation scheme specifically targeting cross-subtopic event coreference. To overcome the aforementioned issue of sparsity, we propose to annotate event mentions at the sentence level. By relying on crowdsourcing, our approach can create an event coreference corpus quickly and at a large scale, opposed to previous annotations which spanned several months and required trained expert annotators [CV14; Vos+18].

2. Using our approach, we create the Football Coreference Corpus (FCC) containing 451 news articles annotated with cross-subtopic event coreference. We demonstrate high agreement between crowd annotators and NLP experts.

2 Related Work

In 2008, Bejan and Harabagiu [BH08] introduced the EventCorefBank as a dataset for CDCR which was later extended by Lee et al. [Lee+12]. Another augmentation effort by Cybulska and Vossen [CV14] then led to the ECB+ dataset. With the goal of improving lexical diversity, they added a second set of documents focusing on a different seminal event for each topic, thereby introducing the notion of *subtopics*. The augmentation effort of adding 502 documents took four months with two expert annotators. Despite the addition of further subtopics, the lexical variety is still low enough that a lemma-matching baseline is a strong contender [Upa+16].

As a consequence, Vossen et al. [Vos+18] developed the Gun Violence Corpus, a corpus consisting of a single topic and 241 subtopics to achieve high lexical diversity. The authors propose the *data-to-text* annotation methodology in which coreference links are established by linking event mention spans in multiple documents to

the same entry in a knowledge base. The corpus was annotated by two annotators over the course of six weeks.

3 Annotation Principles

This section explains the principles of our annotation approach. The application of these principles for annotating the FCC follows in Section 4.

Annotation Unit. As mentioned previously, cross-subtopic event coreference links occur sparsely throughout a set of documents. We therefore propose to annotate event mentions entirely at the sentence level. In doing so, we trade mention span granularity for an increased density of the coreference phenomenon we seek to annotate, which in turn ensures an affordable annotation. Another implication is that the annotation of sentences becomes a multi-label task, because a single sentence can mention multiple events. The definition of what triggers an event mention within a sentence remains a separate question and is unaffected by the change of the annotation unit. A sentence mentioning an event may depend on participants, time or location mentioned elsewhere in the document. Therefore, annotating and resolving mentions requires the full document for context.

Annotation Objective. We frame the annotation task as a variation of event linking. Given a query sentence, several preceding sentences for context and a predefined set of events, annotators need to select the subset of events which are explicitly mentioned in the query sentence. In the end, all sentences linked to the same event e will form an event coreference cluster corresponding to e. Note that the set of events is mainly an aid for annotation and needs not to be provided to coreference resolution systems at training or test time.

Conceptually, the only assets required in our approach are a set of events and a loosely related set of documents. In domains where the set of to-be-annotated events is known prior to the annotation, such a set of events either already exists or is easy to construct. Furthermore, given the complexity of annotating cross-document event coreference, the opportunity of scaling up the annotation far outweighs the comparably small effort of defining such a set.

Past work on event coreference annotation [BH08; Hov+13b; Vos+18] demonstrated that structuring events in a hierarchical fashion offers benefits, for example by separating events that are distinct but conceptually related, which leads to more precise annotations if the hierarchy is relayed to human annotators. We therefore impose a hierarchical structure on the set of to-be-annotated events via the subevent relation. We follow the definitions of Hovy et al. [Hov+13b] and define that an event e_1 is considered a subevent of another event e_2 if its respective action, participants, time and location are subsets of the corresponding properties of e_2. Note that the annotation of a list of unrelated events remains possible with our proposed approach, since a list of events can be reformulated as a flat hierarchy.

Annotation Workflow. For each document, we provide annotators with one predefined event hierarchy. Given a query sentence, we first ask annotators via binary yes/no question whether the sentence mentions events from the event hierarchy. If it does, annotators are supposed to select the subset of events from the hierarchy which are mentioned in the sentence explicitly or by a subevent.

Annotation Aggregation and Agreement. We aggregate the annotations of the binary question into a gold standard using MACE [Hov+13a]. We obtain a multi-label annotation for each sentence and each annotator for the event linking annotations. Based on this information, we need to find the gold set of events E^* which is most representative of all annotations.

Identifying this set is challenging because one needs to distinguish between cases where high variance in the annotations stems from disagreeing annotators or from a genuine case of a sentence deserving multiple labels. Conceptually, any mention of an event implicitly also functions as a mention for all its *super events* (i.e. its ancestors in the event hierarchy). We therefore require that no two events standing in an ancestor relation to each other in the event hierarchy may be present in the gold set of events of a sentence. We framed the search for the optimal set E^* as a constraint optimization problem based on this condition.

We compute inter-annotator agreement using Krippendorff's Alpha [Kri04]. Because our labels are sets of events, we follow the recommendations of Artstein and Poesio [AP08] and use the Jaccard index [Jac12] as a distance metric.

4 A Corpus of Sports News Reports

We applied our annotation approach to create the Football Coreference Corpus (FCC), the first corpus which focuses on cross-subtopic event coreference relations. Due to the sparsity of this event coreference phenomenon, one needs to draw a large sample of articles from the same topic to obtain sufficiently many and sufficiently large coreference clusters. We therefore decided to annotate documents from the sports domain, specifically

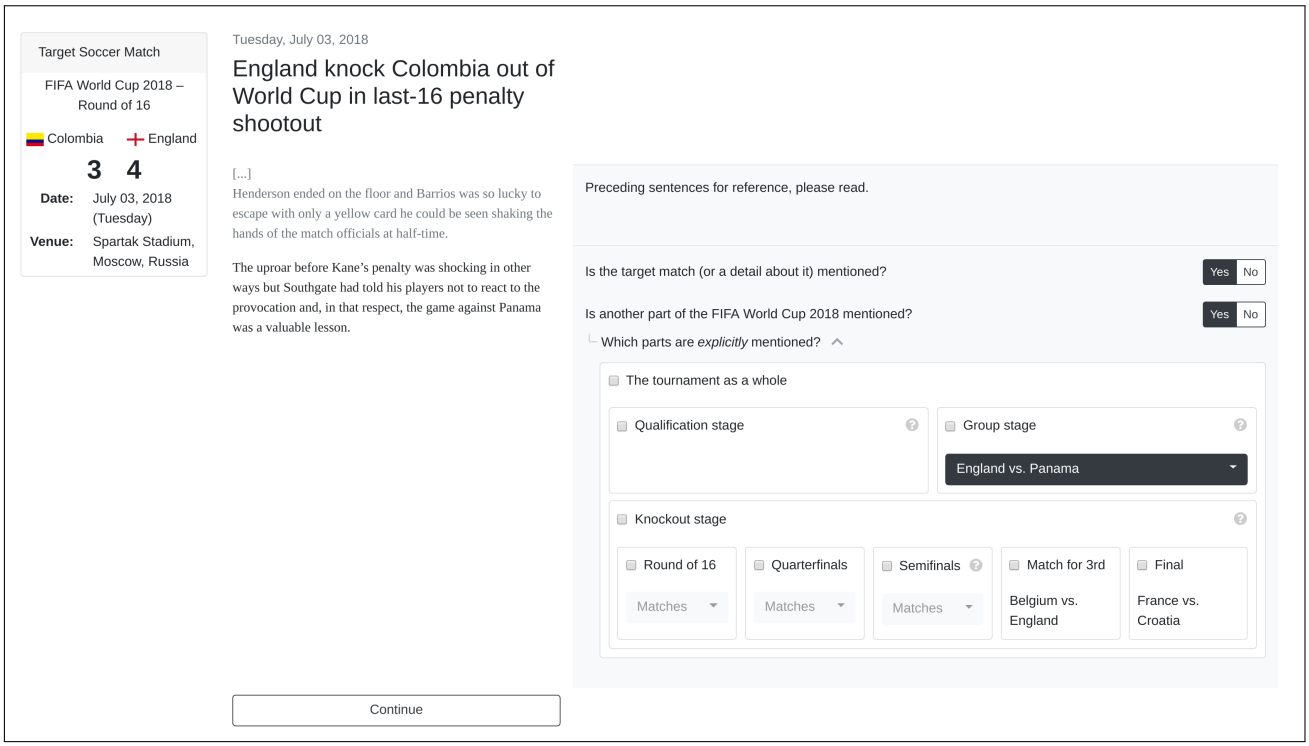


Figure 2: Annotation interface

match reports and related articles about football (soccer) tournaments. As we will explain in the following, our annotation approach can be implemented without requiring domain-specific annotation guidelines which boosts its generalizability.

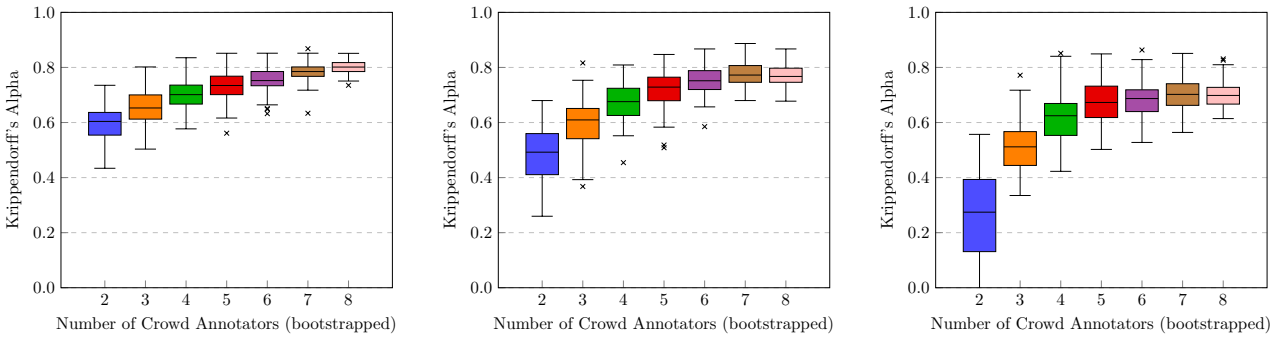
4.1 Data and Methodology

We annotated news articles for five tournaments (FIFA World Cups 2010, 2014, 2018 and UEFA Euro 2012, 2016) independently. Articles were obtained through the Google Custom Search API. Article contents were extracted from the surrounding webpages, followed by a sentence splitting step using NLTK [BKL09; KS06]. We set up a separate event hierarchy for each tournament. We manually determined the seminal event of each article, thereby connecting articles to the event hierarchy and establishing the subtopics of the corpus.

The annotation effort was distributed to crowdworkers on the Amazon Mechanical Turk platform. We developed an annotation interface implementing the approach outlined in Section 3 (see figure 2). Apart from the query sentence, crowdworkers were additionally shown several preceding sentences, the article headline and its publication date to establish the document context. If the seminal event of a document was part of the event hierarchy, we additionally displayed the properties of this event (action, participants, location, date).We kept textual explanations to a minimum and instead primed annotators through a series of examples on how to complete the assignments correctly. In particular, we did not specify rules on what defines an event mention but relied on the annotators' commonsense to make this decision. The only domain-dependent elements in our annotation setup are the examples (which were chosen from the same domain as the input documents) and minor help texts explaining the process of football tournaments.

4.2 Validity

In order to ensure the validity of our study and to determine the minimum number of crowd annotators that provide sufficient quality, we compared crowdworker annotations to those of experts. 121 sentences from three randomly picked documents were independently annotated by 8–9 crowdworkers and two NLP experts. Between the two experts, an agreement of 0.76 Krippendorff's Alpha was reached. We manually created a gold standard from the expert annotations. We bootstrapped the crowdworker annotations to simulate annotation studies with 2 to 8 crowdworkers per instance. Each bootstrapped set of crowdworker annotations was aggregated and



(a) Seminal event y/n (b) Hierarchy event y/n (c) Set of mentioned events

Figure 3: Agreement between experts and different numbers of crowd annotators for each task. Each distribution is the result of 100 bootstrapped runs.

Table 1: Comparison between FCC (our dataset), ECB+ and GVC. *We found 13 annotated cross-subtopic clusters of which 3 stem from annotation errors.

	ECB+ [CV14]	GVC [Vos+18]	FCC (ours)
Unit of annotation	token	token	sentence
Annotators	2 experts	2 experts	crowd
Annotation duration	4 months	6 weeks	3 weeks
Documents	982	510	451
Sentences	16 314	9782	15 529
Topics	43	1	1
Subtopics per topic	2	241	95
Events	8911	1411	217
├ Singletons	8010	365	48
├ Within-doc chains	20	301	10
└ Cross-doc clusters	881	745	159
└ **Cross-subtopic clusters**	10*	0	142
Mentions	15 003	7298	2618

compared to the experts' work. Figure 3 shows the resulting agreement distributions for the three tasks in our annotation scheme (mention detection for seminal events, mention detection for other events, annotating the set of referenced events). With five crowd annotators, we reach a mean Krippendorff's Alpha of 0.734, 0.719 and 0.677, allowing tentative conclusions [Car96]. This constitutes a suitable tradeoff between quality and costs, leading us to annotate the main portion of the dataset with five crowd annotators.

4.3 Execution and Analysis

Using time measurements from pre-studies as a reference point, we paid annotators according to the US minimum wage of $7.25/hour. Overall, 451 documents were annotated over the course of three weeks, amounting to $2800 in total.

Table 1 shows the properties of our dataset alongside a comparison to existing CDCR datasets. The contrast in the overall number of annotated events is a result of different annotation strategies: In ECB+, all mentions of a document's seminal event as well as any other events mentioned in the same sentence were annotated. For the Gun Violence Corpus (GVC), only mentions of a given seminal event and a predefined set of its subevents were annotated. In our dataset, 311 coarse-grained events were available for annotation across all event hierarchies. In the end, 217 of these events were annotated by crowd annotators. Given their higher granularity compared to those annotated in ECB+ and the GVC, these events are less frequent by nature. Most notably however, our proposed annotation scheme resulted in a dataset with a large number of cross-subtopic event coreference clusters. While the annotation of this type of coreference is technically possible with traditional token-level annotation schemes, ours is markedly faster and does not require complex or domain-dependent annotation guidelines or trained annotators, which to the best of our knowledge makes it the first scalable technique for

Table 2: Crowdsourced data examples. Each sentence is taken from different documents belonging to different subtopics. There are two cross-subtopic event coreference clusters: sentences [1,2,3] and [2,3]. Note the annotation of a future event in sent. 1 and of an expression involving a quantifier in sent. 3.

	Subtopic	Sentence	Annotated Events
1	CRO vs. DEN R16_4	*Awaiting Saturday, a quarterfinal date with the Russian hosts in Sochi.*	CRO vs. RUS QF_3
2	CRO vs. ENG SF_2	*But it's fair to say they're taking the harder route, having followed up back-to-back penalty shootout wins over Denmark and Russia by coming from behind to make their way past England.*	CRO vs. RUS QF_3 CRO vs. DEN R16_4
3	FRA vs. CRO Final	*Croatia had played extra time in each of its three previous matches but showed no signs of fatigue early in the final.*	CRO vs. RUS QF_3 CRO vs. DEN R16_4 CRO vs. ENG SF_2

annotating cross-subtopic event coreference links.

Table 2 shows exemplary results from our annotation. A strong point of our approach relying on commonsense is that crowdworkers also linked future events and expressions with quantifiers triggering multiple events ("three previous matches") without us having to provide detailed annotation guidelines on how to handle these cases. We manually analyzed a number of annotated documents. In some cases, multiple events are mentioned via non-countable quantifiers ("few", "every", "more than two", etc.). This caused annotators to agree on the presence of a mention but caused disagreement for the linking step. In case annotators reach no consensus with respect to the set of mentioned events, the τ parameter in our aggregation returns an empty gold set of events. Some sentences cause disagreement because an event mention leaves room for interpretation as to which event is being referenced. For example, in the sentence "World Cup 2018: France beat Uruguay 2-0 to reach semi-final" it is unclear whether the semifinal *match* or the superevent semifinal *stage* is mentioned.

5 Conclusion

We are, to the best of our knowledge, the first to tackle cross-subtopic event coreference, a salient but rarely occurring coreference phenomenon which is underrepresented in other datasets. To capture these links affordably and with sufficient density in text, we developed a novel sentence-level crowdsourcing annotation scheme, which produces reliable results when compared to NLP experts. We created the Football Coreference Corpus (FCC), the first CDCR corpus specifically targeting cross-subtopic event coreference which consists of 451 football news reports. Our work offers several possibilities for follow-up work: Since our proposed annotation scheme does not require domain-specific annotation guidelines, future work may add further topics with relative ease.

Acknowledgments

The authors would like to thank the anonymous reviewers for their helpful insights. This work was supported by the German Research Foundation under grant №GU 798/17-1.

References

[AP08] Ron Artstein and Massimo Poesio. "Inter-coder agreement for computational linguistics". In: *Computational Linguistics* 34.4 (2008), pp. 555–596.

[BH08] Cosmin Bejan and Sanda Harabagiu. "A Linguistic Resource for Discovering Event Structures and Resolving Event Coreference". In: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco: European Language Resources Association (ELRA), May 2008. ISBN: 2-9517408-4-0.

[BH14] Cosmin Adrian Bejan and Sanda Harabagiu. "Unsupervised event coreference resolution". In: *Computational Linguistics* 40.2 (2014), pp. 311–347. DOI: 10.1162/COLI_a_00174.

[BKL09] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, 2009.

[Car96] Jean Carletta. "Assessing Agreement on Classification Tasks: The Kappa Statistic". In: *Computational Linguistics* 22.2 (1996), pp. 249–254. URL: https://www.aclweb.org/anthology/J96-2004.

[CV14] Agata Cybulska and Piek Vossen. "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)*. Reykjavik, Iceland: European Language Resources Association (ELRA), May 2014. ISBN: 978-2-9517408-8-4.

[Hov+13a] Dirk Hovy et al. "Learning Whom to Trust with MACE". In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Atlanta, Georgia: ACL, June 2013, pp. 1120–1130. URL: http://www.aclweb.org/anthology/N13-1132.

[Hov+13b] Eduard Hovy et al. "Events are Not Simple: Identity, Non-Identity, and Quasi-Identity". In: *Workshop on Events: Definition, Detection, Coreference, and Representation*. Atlanta, Georgia: ACL, June 2013, pp. 21–28. URL: http://www.aclweb.org/anthology/W13-1203.

[Jac12] Paul Jaccard. "The distribution of the flora in the alpine zone". In: *New phytologist* 11.2 (1912), pp. 37–50.

[Kri04] Klaus Krippendorff. *Content Analysis, an Introduction to Its Methodology, 2nd Edition*. Thousand Oaks, CA: Sage Publications, 2004.

[KS06] Tibor Kiss and Jan Strunk. "Unsupervised Multilingual Sentence Boundary Detection". In: *Computational Linguistics* 32.4 (2006), pp. 485–525. DOI: 10.1162/coli.2006.32.4.485. eprint: https://doi.org/10.1162/coli.2006.32.4.485. URL: https://doi.org/10.1162/coli.2006.32.4.485.

[Lee+12] Heeyoung Lee et al. "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*. Jeju Island, Korea: ACL, 2012, pp. 489–500. URL: http://www.aclweb.org/anthology/D12-1045.

[Min+15] Anne-Lyse Minard et al. "SemEval-2015 Task 4: TimeLine: Cross-Document Event Ordering". In: *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Denver, Colorado: ACL, June 2015, pp. 778–786. DOI: 10.18653/v1/S15-2132. URL: https://www.aclweb.org/anthology/S15-2132.

[Mor99] Thomas S. Morton. "Using Coreference for Question Answering". In: *Proceedings of the Eighth Text REtrieval Conference (TREC 8)*. 1999, pp. 685–688. URL: https://www.aclweb.org/anthology/W99-0212.

[PIV18] Marten Postma, Filip Ilievski, and Piek Vossen. "SemEval-2018 Task 5: Counting Events and Participants in the Long Tail". In: *Proceedings of The 12th International Workshop on Semantic Evaluation*. New Orleans, Louisiana: ACL, June 2018, pp. 70–80. URL: http://www.aclweb.org/anthology/S18-1009.

[Upa+16] Shyam Upadhyay et al. "Revisiting the evaluation for cross document event coreference". In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. 2016, pp. 1949–1958.

[Vos+18] Piek Vossen et al. "Don't Annotate, but Validate: a Data-to-Text Method for Capturing Event Data". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), May 2018. ISBN: 979-10-95546-00-9.