Evaluating Zero-Shot Event Structures: Recommendations for Automatic Content Extraction (ACE) Annotations

Erica Cai **Brendan O'Connor**
University of Massachusetts Amherst
{ecai,brenocon}@cs.umass.edu

Abstract

Zero-shot event extraction (EE) methods infer richly structured event records from text, based only on a minimal user specification and no training examples, which enables flexibility in exploring and developing applications. Most event extraction research uses the Automatic Content Extraction (ACE) annotated dataset to evaluate *supervised* EE methods, but can it be used to evaluate *zero-shot* and other low-supervision EE? We describe ACE's event structures and identify significant ambiguities and issues in current evaluation practice, including (1) coreferent argument mentions, (2) conflicting argument head conventions, and (3) ignorance of modality and event class details. By sometimes mishandling these subtleties, current work may dramatically understate the actual performance of zero-shot and other low-supervision EE, considering up to 32% of correctly identified arguments and 25% of correctly ignored event mentions as false negatives. For each issue, we propose recommendations for future evaluations so the research community can better utilize ACE as an event evaluation resource.

1 Introduction

Zero-shot event extraction (EE) methods infer richly structured instances of action or relationship occurrences from unstructured text data, based on a user-supplied natural language specification of the desired event—without annotated training examples (Du and Cardie, 2020; Liu et al., 2020; Li et al., 2021; Lyu et al., 2021). The extracted structure is useful for many applications such as analyzing interactions between entities and performing more intelligent question answering (Gao et al., 2016; Liu et al., 2017a; Cao et al., 2020; Li et al., 2020b), and the low resources required by zero-shot EE methods further this practical advantage. We refer to the structure as an event, where each event could have an arbitrary structure as needed. Each structure contains information such as the participants involved, content, and location of the event.

To evaluate *supervised* EE methods, many works use the Automatic Content Extraction (ACE) dataset—specifically, the Linguistic Data Consortium's *ACE 2005 Multilingual Training Corpus* (Doddington et al., 2004),[1] which includes English, Chinese, and Arabic documents and resulted from the U.S. federal government's ACE program.[2] The ACE dataset stores information about entities, relations, and events from 598 (for English) documents in a rich structure; our focus is mostly on its events. ACE is frequently used for event extraction modeling and evaluation, and is often claimed to be the most widely used such dataset (§3). While there are many somewhat similar structured semantic datasets, ACE still shines in having whole-document annotations (contra FrameNet; Baker et al., 2003; Baker and Sato, 2003; Fillmore et al., 2003), realistically non-lexical-specific event classes (contra PropBank (Palmer et al., 2005), OntoNotes (Weischedel et al., 2017), and Semantic Dependencies (Oepen et al., 2014)), event modality (contra PB, ON, SD), English data (contra Entities, Relations, and Events (ERE)),[3] and specification of event arguments (contra Richer Event Description (RED); O'Gorman et al., 2016) that are simultaneously represented both as text spans

[1] https://catalog.ldc.upenn.edu/LDC2006T06 https://doi.org/10.35111/mwxc-vh88

[2] https://www.ldc.upenn.edu/collaborations/past-projects/ace, http://web.archive.org/web/20080303183132/https://www.nist.gov/speech/tests/ace/. A separate evaluation dataset was not released publicly (Haghighi and Klein, 2009, footnote 7); we follow the convention of subsequent research of referring to the public release LDC2006T06 as "the ACE dataset" or simply "ACE," despite "training" in its title.

[3] Song et al. (2015) promise an LDC release of their ERE annotations while Aguilar et al. (2014) analyzes ERE's guidelines—both with English examples—but LDC's catalog suggests only a Chinese corpus was ever released (LDC2020T19). Li et al. (2020b) reports ERE English results, presumably from a proprietary dataset.

(contra Abstract Meaning Representation (AMR); Banarescu et al., 2013), and discourse-level entities[4] (§2). While ACE does not include RED's interesting causal and bridging event-event relations (see also Hovy et al., 2013), its core tasks related to entities and event arguments have important applications and are far from solved.

We investigate using the ACE dataset to evaluate *zero-shot* and other low-supervision EE methods, which are more real-world relevant than highly-supervised EE methods for requiring few if any annotations, but which may face certain evaluation challenges more severely.[5] First, we identify issues related to how evaluations extract gold event argument annotations from ACE and to the possibly clashing use case of a zero-shot EE method versus the annotations in ACE. Evaluation of zero-shot EE methods is particularly sensitive to these issues since they lack knowledge of (sometimes arbitrary) details in ACE event structures that are implicit in training examples—and their ignorance of them may be correct for many applications. Therefore, we present guidelines and methods to overcome these issues in English, which could in theory be adaptable to other languages, and quantify their potential impact.[6]

2 Structure of Events and Entities in ACE

The Automatic Content Extraction (ACE) dataset stores annotations for entity, relation, and event structures for news, conversations, blog, and transcript textual data. We focus on the ACE event extraction task (Ahn, 2006), which takes a sentence as input and outputs a set of event tuples, which we attempt to precisely specify.

Events (Figure 1). Every event takes one of 33 discrete event classes $t \in \mathcal{T}$,[7] each of which has a discrete set of (typically, 1–6) roles $\mathcal{R}_t$ which its arguments can take. An event tuple has the form $\langle t, g, \{a_1..a_n\} \rangle$ where

1. $t \in \mathcal{T}$ is the event class.
2. g is the span[8] of the **event trigger**, a word that identifies or represents the event class.
3. $\{a_1..a_n\}$ is a (possibly empty) set of **event arguments** explicitly mentioned in the sentence, each with $a_i = \langle a_i^{(r)}, a_i^{(s)} \rangle$: the **role** $a^{(r)} \in \mathcal{R}_t$, and argument span $a^{(s)}$.[9]

The full *event extraction* task is to output some number of event tuples from the sentence; research often examines subtasks to identify various subsets of $t, g, a^{(r)}, a^{(s)}$, such as *event trigger classification* or *event detection* (just (g, t)). Finally, the tuple has several additional semantic tags such as modality and tense (§4.3).

Entities (Figure 2). An event argument $a^{(s)}$ may also be a *mention* of an *entity*, a document-level object with its own type information and one or more coreferential mention spans throughout a document. For an argument span $a^{(s)}$, let $\mathcal{C}(a^{(s)})$ refer to the set of all its coreferential mentions.[10] Additionally, ACE's `<entity>` data structure defines for each mention a **head** span (§4.2).

In the following example from ACE, a killing (LIFE.DIE) event has agent $a^{(s)}$="Iraq's Mukhabarat" (Figure 1); when cross-referencing the entity information $\mathcal{C}(a^{(s)})$, it turns out this argument is coreferentially mentioned three times in the sentence (Figure 2).

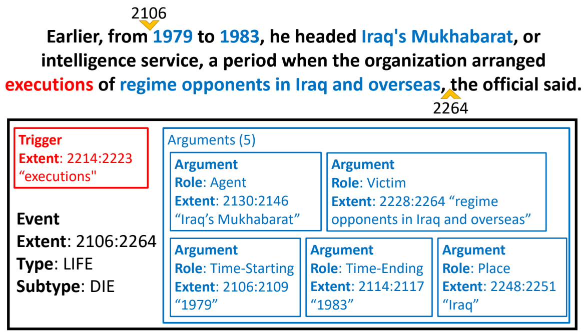

Figure 1: Event tuple in a sentence from ACE document `APW_ENG_20030417.0555`.

[4] ON 5.0 does include joint coreference and PropBank annotations, albeit for only 18% of its English documents.

[5] This work is motivated by the zero-shot setting (no annotated data, but with user specification of desired event class), but could apply to other low-supervision settings including unsupervised event extraction (e.g. Chambers (2013): induces event classes with no annotations or user guidance) and few-shot EE (small number of annotated examples). In all low supervision settings, we do not expect a model to learn superficial annotation quirks, which we believe is more real-world relevant than high-supervision settings.

[6] While we do not contribute new evaluation software, we make available code to reproduce this paper's analyses at: https://github.com/ec769/ZS-evalanalysis-ACE.

[7] ACE defines 7 event *types* and 33 *subtypes*; we do not focus on this hierarchical structure.

[8] In ACE and both figures, called an *extent*, which is a character span (start and end positions) in the text. We use *span* and *extent* interchangeably.

[9] Each argument may be a time, value (i.e. quantity), or entity (i.e. person, place or thing); most works discussed in §3 consider only entities as arguments, which we follow.

[10] Our discussion conflates an argument span with the entity's mention span at the same location; ACE technically defines them separately, and sometimes the argument span can be longer. However, they always have the same head.

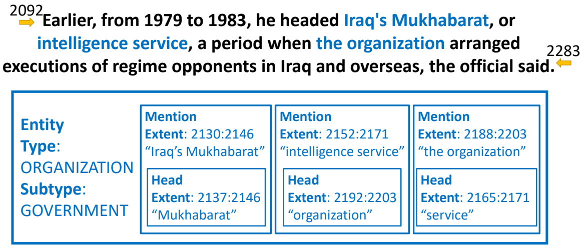

Figure 2: Three coreferential mentions $\mathcal{C}$ of the Figure 1's AGENT; see §4.1.

3 Review of Using ACE to Evaluate EE

We reviewed 38 papers published from 2008 through 2022, cited in Li et al. (2022)'s survey of deep learning methods for event extraction, to examine how they use ACE to evaluate EE tasks (Ji and Grishman, 2008; Liao and Grishman, 2010; Hong et al., 2011; Li et al., 2013; Nguyen and Grishman, 2015; Chen et al., 2015; Nguyen et al., 2016; Yang and Mitchell, 2016; Nguyen and Grishman, 2016; Feng et al., 2016; Liu et al., 2016; Huang et al., 2016; Sha et al., 2016; Chen et al., 2017; Liu et al., 2017b; Zhao et al., 2018; Zeng et al., 2018; Hong et al., 2018; Liu et al., 2018; Huang et al., 2018; Liu et al., 2019; Zhang et al., 2019b; Wang et al., 2019; Zhang et al., 2019a; Yang et al., 2019; Nguyen and Nguyen, 2019; Wadden et al., 2019; Chen et al., 2020; Du and Cardie, 2020; Liu et al., 2020; Li et al., 2020a; Lin et al., 2020; Li et al., 2021; Ahmad et al., 2021; Zhou et al., 2021; Wang et al., 2021; Lu et al., 2021; Lyu et al., 2021). Several state that ACE is the most popular dataset for evaluating EE methods (Li et al., 2022; Zhang et al., 2019b; Wang et al., 2019). While the ACE data release does not define a split, these papers, especially after 2011, settled on a shared train/development/test split (§A.6).

When considering event trigger and event argument identification, all papers require matching the gold standard's extent to be considered correct. For arguments, which are usually multiple tokens long, some works require matching the full argument extent $a^{(s)}$ while others only use its head extent. (Additional details in §A.6.)

The works that we analyzed identify several challenges with using ACE. Some event subtypes are very sparse; almost 60% of event types have fewer than 100 labeled samples, while three event types each have fewer than ten out of the 5042 samples over all English documents and 33 event classes (Chen et al., 2017; Liu et al., 2017b, 2018). Second, the manually specified event schemas in ACE are hard to generalize to different domains, such as the WEAPON argument role (Huang et al., 2016). Third, Ji and Grishman (2008) find that human annotators achieve only about 73% of the F1 score on the argument and trigger identification task and annotation quality continues to be questioned in debates about annotation guidelines (Lin et al., 2020). In any case, ACE remains a widely used dataset for evaluation.

4 Recommendations for Using ACE to Evaluate Zero-Shot and Other Low-Supervision EE Methods

Recommendation 1: Coreference Invariant Argument Matching. *To evaluate correctness of event arguments using ACE, allow a match to* ***any*** *coreferent mention of the argument ($c \in \mathcal{C}(a^{(s)})$), not just the one mention in* <event_argument_mention> *($a^{(s)}$).* This (we believe) erroneous practice is widespread, and may consider up to 32% of correctly identified entity-type arguments as incorrect.

Problem. Although ACE stores event triggers and types as part of an *event mention*, it stores event arguments as part of both *event mentions* and *entity, time,* or *value mentions*. The *event mention argument* stores *one* reference to the argument ($a^{(s)}$), even if multiple references exist ($\mathcal{C}(a^{(s)})$). Low supervision EE methods can not learn a training set's potentially superficial convention for which of multiple references to specify.

Issues in the Literature. Alarmingly, although ACE stores multiple gold references as entity mentions, they are often not used. We find that a number of recent works, especially on zero-shot EE, that ignore them. Wadden et al. (2019)'s preprocessing code, which was used in several later works (Du and Cardie, 2020; Lin et al., 2020; Li et al., 2021; Lu et al., 2021; Lyu et al., 2021), does not gather multiple references to $a^{(s)}$ in an event tuple. While an unofficial update includes entity information, we identify further difficulties in §A.3. Independently, Zeng et al. (2018) acknowledge not applying coreference resolution, which contributes to a higher argument identification task error rate. While we acknowledge that whether to model coreference is a complex question, using gold standard coreference information at *evaluation* time is an independent issue and ought to be mandatory, for any modeling approach. Even for a purely extent prediction system, gold-standard coreference is necessary for correct evaluation.

# Refs Per Arg	1	2	3	4	5	6
Excl. Pronoun	85.36	12.11	2.25	0.25	0.03	0
Incl. Pronoun	68.28	21.45	7.34	2.37	0.43	0.13

Table 1: The percent of arguments with a varying number of references to it ($|\mathcal{C}(a)|$) in the same sentence, excluding duplicates, where pronouns are and are not arguments. (More implementation details in §A.5).

Findings. Table 1 shows that roughly 14.6% of arguments have multiple references within the same sentence when arguments are not pronouns, and roughly 31.7% do otherwise. In the worst case, an evaluation could consider all such arguments, even if correctly identified, as false negatives. Next, we investigate if a pattern for choosing $a^{(s)}$ out of $\mathcal{C}(a^{(s)})$ exist. If multiple references exist, $a^{(s)}$ is the first reference to appear in the sentence 56.3% of the time. If one or more is a named entity, $a^{(s)}$ is, also, 60.7% of times. (More details in §A.5). Given the alarming statistics in Table 1 and a non-observable pattern for choosing $a^{(s)}$ out of $\mathcal{C}(a^{(s)})$, we recommend extracting all possible references to an argument using the `<entity>` object, instead of only relying on `<event mention argument>`.

Recommendation 2: Dual ACE and Automatic Head Selection. *To evaluate correctness of an event argument using ACE, in addition to comparing its head against the head provided by ACE, compare its head against the one selected by a Universal Dependency-based parser. We find that 8.1% of heads that the English portion of ACE identifies are not consistent with a Universal Dependency SpaCy3 parser-based head finder (more details in §A.4.1).*

Problem and Literature. To determine correctness of an event argument, either compare it against $a^{(s)}$ in ACE or its head against $a^{(s)}$'s head. Comparing against the entire $a^{(s)}$ is likely to yield false negatives because ACE argument spans can be very long, including the noun phrase's complements and even elaborate relative clauses (e.g.: "the women from Texas who heinously drowned her five kids, aged 6 months to 7 years, one by one, in her bathtub"). Thus most works we reviewed evaluate argument correctness by comparing its head with the head of potential $a^{(s)}$s. For zero-shot EE methods with no knowledge of argument constitutions, using the head seems especially appropriate.

Method and Findings. We investigate if the head of $a^{(s)}$ that ACE specifies is consistent with the Universal Dependency (UD) (Nivre et al., 2020) definition of head, which we identify from spaCy3's UD parse as the token in the span that is an ancestor to the rest of the span (i.e., the span's subgraph's root); we additionally add a heuristic to address a frequent parse error when the noun phrase head is analyzed as the relative clause's subject, and to extend the head to be multiple tokens (as sometimes occurs in ACE's heads) when the head token is within an spaCy3-identified named entity. The discrepancies in Figure 3 suggest that ACE often does not follow the UD formalism. (Additional algorithmic details and discrepancies in §A.4.)

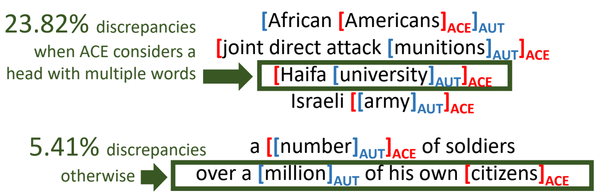

Figure 3: Examples of discrepancies between the head in ACE and the head identified by the UD-based algorithm, and percentages of such discrepancies when ACE considers a multi-word head versus a single-word head. Each line contains an argument extent; the head by ACE is in red brackets and that by UD is in blue.

Next, we explore the feasibility of consistently reconstructing the exact head specified by ACE. Given clear inconsistencies in the way that ACE selects the head in Figure 3 (eg: "Haifa university" and "Israeli army"), we conclude that ACE may not identify the argument head in a systematic or at least easily emulatable way, which may contribute to false negatives. To eliminate the inconsistency issue, we propose to use a UD-based algorithm to select heads from ACE argument extents for matching, *in addition* to the heads specified by ACE. The head from the UD-based parser is not always the most appropriate for a given argument extent (see error analysis of parser behavior in §A.4.1), but our approach does avoid the *inconsistency* issue. While we only applied our UD-based algorithm to English data, this head-matching approach may be adaptable to other languages with available UD parsers.

Recommendation 3: Analyze a Subset of ACE Modalities or Event Classes. *Consider a subset of annotated events as the ground truth event set to improve the evaluation of zero-shot EE methods that target a particular use case; e.g., sociopolitical analysis.*

Problem and Literature. While greater flexibility enables zero-shot EE methods to be more practical, extracting structured data as events without requiring training examples, each practical application has a different objective. For example, social scientists and political forecasters may need to analyze historical events that actually happened in the past (Schrodt et al., 1994; O'Connor et al., 2013; Boschee et al., 2013; Halterman et al., 2021; Hanna, 2017; Hürriyetoğlu et al., 2021; Giorgi et al., 2021; Stoehr et al., 2021), such as in the widely-used ICEWS automatically generated events dataset (Boschee et al., 2017). However, in other applications such as those on opinion or sentiment tasks (Swamy et al., 2017), the aim of zero-shot EE methods may be benefited by hypothetical events.

Many aspects of modality have been explored in computational modeling, such as temporal semantics (Timebank (Pustejovsky et al., 2003)), factual versus uncertain or hypothetical status (Factbank (Saurí and Pustejovsky, 2009), Pragbank (de Marneffe et al., 2012), (Diab et al., 2009; Prabhakaran et al., 2015; Stanovsky et al., 2017; Rudinger et al., 2018; Yao et al., 2021; Lee et al., 2015)), and in literary domains (Litbank (Bamman et al., 2019, 2020)). ACE includes a simple modality label for each event instance as either `ASSERTED` to indicate an event instance that was referred to as a real occurrence, or `OTHER` for all others: non-grounded beliefs (e.g. rumors), hypotheticals, commands, threats, proposals, desires, promises, etc. In fact, for 25% of event instances in ACE, the modality tag label is `OTHER`. Yet, the 38 works that we explored in §3 which use ACE to evaluate EE methods do not include modality as part of the task definition. We propose that future work could better use ACE by predicting or analyzing subsets of modalities to more clearly support downstream applications.

Finally, modality is important since it may also interact with modeling (Cai and O'Connor, 2023). Zero-shot EE methods involving question-answering (QA) or text entailment (TE) models (Lyu et al., 2021), may enforce modality restrictions through the language in the query. For example, the past tense question "did the police arrest someone?" (Halterman et al., 2021) asks for a reported occurrence that the police are arresting or have arrested someone, but not an intended or hypothetical arrest. Whether this matches user intent, and whether models respect or ignore the query's modality restrictions, are important avenues for future work; ACE data can aid such analysis.

5 Conclusion

We explore how to use ACE, which is a gold standard dataset containing annotations of events from diverse text data in a rich structure, to evaluate zero-shot and other low-supervision EE methods by identifying issues that may more severely affect their evaluation. We particularly find difficulties with evaluating spans of events due to a lack of training data for zero-shot and low-supervision EE methods to learn superficial annotation quirks from. However, we present methods to overcome these issues and demonstrate them on the English portion of ACE, noting that in principle they may be adaptable to any language. Ultimately, we advocate for using ACE to evaluate zero-shot and other low-supervision EE methods after addressing the issues, and discuss the potential for using ACE in smarter ways to evaluate different types of EE methods in the future.

Acknowledgments

We thank the UMass NLP group and anonymous reviewers for feedback. This work was supported by NSF CAREER 1845576. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

1991. *Third Message Uunderstanding Conference (MUC-3): Proceedings of a Conference Held in San Diego, California, May 21-23, 1991.*

Jacqueline Aguilar, Charley Beller, Paul McNamee, Benjamin Van Durme, Stephanie Strassel, Zhiyi Song, and Joe Ellis. 2014. A comparison of the events and relations across ACE, ERE, TAC-KBP, and FrameNet annotation standards. In *Proceedings of the Second Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 45–53, Baltimore, Maryland, USA. Association for Computational Linguistics.

Wasi Ahmad, Nanyun Peng, and Kai-Wei Chang. 2021. Gate: Graph attention transformer encoder for cross-lingual relation and event extraction. In *The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21).*

David Ahn. 2006. The stages of event extraction. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, pages 1–8, Sydney, Australia. Association for Computational Linguistics.

Collin Baker, Charles Fillmore, and Beau Cronin. 2003. The structure of the framenet database. *International Journal of Lexicography*, 16:281–296.

Collin F. Baker and Hiroaki Sato. 2003. The FrameNet data and software. In *The Companion Volume to the Proceedings of 41st Annual Meeting of the Association for Computational Linguistics*, pages 161–164, Sapporo, Japan. Association for Computational Linguistics.

David Bamman, Olivia Lewke, and Anya Mansoor. 2020. An annotated dataset of coreference in English literature. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 44–54, Marseille, France. European Language Resources Association.

David Bamman, Sejal Popat, and Sheng Shen. 2019. An annotated dataset of literary entities. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2138–2144, Minneapolis, Minnesota. Association for Computational Linguistics.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Elizabeth Boschee, Jennifer Lautenschlager, Sean O'Brien, Steve Shellman, James Starz, and Michael Ward. 2017. ICEWS coded event data.

Elizabeth Boschee, Premkumar Natarajan, and Ralph Weischedel. 2013. Automatic extraction of events from open source text for predictive forecasting. *Handbook of Computational Approaches to Counterterrorism*, page 51.

Erica Cai and Brendan O'Connor. 2023. A monte carlo language model pipeline for zero-shot sociopolitical event extraction.

Qingqing Cao, Harsh Trivedi, Aruna Balasubramanian, and Niranjan Balasubramanian. 2020. DeFormer: Decomposing pre-trained transformers for faster question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4497, Online. Association for Computational Linguistics.

Nathanael Chambers. 2013. Event schema induction with a probabilistic entity-driven model. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Seattle, Washington, USA. Association for Computational Linguistics.

Yubo Chen, Shulin Liu, Xiang Zhang, Kang Liu, and Jun Zhao. 2017. Automatically labeled data generation for large scale event extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 409–419, Vancouver, Canada. Association for Computational Linguistics.

Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 167–176, Beijing, China. Association for Computational Linguistics.

Yunmo Chen, Tongfei Chen, Seth Ebner, Aaron Steven White, and Benjamin Van Durme. 2020. Reading the manual: Event extraction as definition comprehension. In *Proceedings of the Fourth Workshop on Structured Prediction for NLP*, pages 74–83, Online. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Christopher D. Manning, and Christopher Potts. 2012. Did it happen? The pragmatic complexity of veridicality assessment. *Computational Linguistics*, 38(2):301–333.

Mona Diab, Lori Levin, Teruko Mitamura, Owen Rambow, Vinodkumar Prabhakaran, and Weiwei Guo. 2009. Committed belief annotation and tagging. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, pages 68–73, Suntec, Singapore. Association for Computational Linguistics.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Xinya Du and Claire Cardie. 2020. Event extraction by answering (almost) natural questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 671–683, Online. Association for Computational Linguistics.

Xiaocheng Feng, Lifu Huang, Duyu Tang, Heng Ji, Bing Qin, and Ting Liu. 2016. A language-independent neural network for event detection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 66–71, Berlin, Germany. Association for Computational Linguistics.

Charles J. Fillmore, Christopher R. Johnson, and Miriam R. L. Petruck. 2003. Background to framenet. *International Journal of Lexicography*, 16:235–250.

Li Gao, Jia Wu, Zhi Qiao, Chuan Zhou, Hong Yang, and Yue Hu. 2016. Collaborative social group influence for event recommendation. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, CIKM '16, page 1941–1944, New York, NY, USA. Association for Computing Machinery.

Salvatore Giorgi, Vanni Zavarella, Hristo Tanev, Nicolas Stefanovitch, Sy Hwang, Hansi Hettiarachchi, Tharindu Ranasinghe, Vivek Kalyan, Paul Tan, Shaun Tan, Martin Andrews, Tiancheng Hu, Niklas Stoehr, Francesco Ignazio Re, Daniel Vegh, Dennis Atzenhofer, Brenda Curtis, and Ali Hürriyetoğlu. 2021. Discovering black lives matter events in the United States: Shared task 3, CASE 2021. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 218–227, Online. Association for Computational Linguistics.

Aria Haghighi and Dan Klein. 2009. Simple coreference resolution with rich syntactic and semantic features. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1152–1161, Singapore. Association for Computational Linguistics.

Andrew Halterman, Katherine Keith, Sheikh Sarwar, and Brendan O'Connor. 2021. Corpus-level evaluation for event QA: The IndiaPoliceEvents corpus covering the 2002 Gujarat violence. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4240–4253, Online. Association for Computational Linguistics.

Alex Hanna. 2017. MPEDS: Automating the generation of protest event data.

Yu Hong, Jianfeng Zhang, Bin Ma, Jianmin Yao, Guodong Zhou, and Qiaoming Zhu. 2011. Using cross-entity inference to improve event extraction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1127–1136, Portland, Oregon, USA. Association for Computational Linguistics.

Yu Hong, Wenxuan Zhou, Jingli Zhang, Guodong Zhou, and Qiaoming Zhu. 2018. Self-regulation: Employing a generative adversarial network to improve event detection. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 515–526, Melbourne, Australia. Association for Computational Linguistics.

Eduard Hovy, 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, Georgia. Association for Computational Linguistics.

Lifu Huang, Taylor Cassidy, Xiaocheng Feng, Heng Ji, Clare R. Voss, Jiawei Han, and Avirup Sil. 2016. Liberal event extraction and event schema induction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 258–268, Berlin, Germany. Association for Computational Linguistics.

Lifu Huang, Heng Ji, Kyunghyun Cho, Ido Dagan, Sebastian Riedel, and Clare Voss. 2018. Zero-shot transfer learning for event extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2160–2170, Melbourne, Australia. Association for Computational Linguistics.

Ali Hürriyetoğlu, Osman Mutlu, Erdem Yörük, Farhana Ferdousi Liza, Ritesh Kumar, and Shyam Ratan. 2021. Multilingual protest news detection - shared task 1, CASE 2021. In *Proceedings of the 4th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2021)*, pages 79–91, Online. Association for Computational Linguistics.

Heng Ji and Ralph Grishman. 2008. Refining event extraction through cross-document inference. In *Proceedings of ACL-08: HLT*, pages 254–262, Columbus, Ohio. Association for Computational Linguistics.

Kenton Lee, Yoav Artzi, Yejin Choi, and Luke Zettlemoyer. 2015. Event detection and factuality assessment with non-expert supervision. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1643–1648, Lisbon, Portugal. Association for Computational Linguistics.

Fayuan Li, Weihua Peng, Yuguang Chen, Quan Wang, Lu Pan, Yajuan Lyu, and Yong Zhu. 2020a. Event extraction as multi-turn question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 829–838, Online. Association for Computational Linguistics.

Manling Li, Alireza Zareian, Ying Lin, Xiaoman Pan, Spencer Whitehead, Brian Chen, Bo Wu, Heng Ji, Shih-Fu Chang, Clare Voss, Daniel Napierski, and Marjorie Freedman. 2020b. GAIA: A fine-grained multimedia knowledge extraction system. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 77–86, Online. Association for Computational Linguistics.

Qi Li, Heng Ji, and Liang Huang. 2013. Joint event extraction via structured prediction with global features. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 73–82, Sofia, Bulgaria. Association for Computational Linguistics.

Qian Li, Hao Peng, Jianxin Li, Yiming Hei, Rui Sun, Jiawei Sheng, Shu Guo, Lihong Wang, and Philip S. Yu. 2022. A survey on deep learning event extraction: approaches and applications. *IEEE Transactions on Neural Networks and Learning Systems*.

Sha Li, Heng Ji, and Jiawei Han. 2021. Document-level event argument extraction by conditional generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 894–908, Online. Association for Computational Linguistics.

Shasha Liao and Ralph Grishman. 2010. Using document level cross-event inference to improve event extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 789–797, Uppsala, Sweden. Association for Computational Linguistics.

Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. 2020. A joint neural model for information extraction with global features. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7999–8009, Online. Association for Computational Linguistics.

Chun-Yi Liu, Chuan Zhou, Jia Wu, Hongtao Xie, Yue Hu, and Li Guo. 2017a. Cpmf: A collective pairwise matrix factorization model for upcoming event recommendation. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 1532–1539.

Jian Liu, Yubo Chen, Kang Liu, Wei Bi, and Xiaojiang Liu. 2020. Event extraction as machine reading comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1641–1651, Online. Association for Computational Linguistics.

Jian Liu, Yubo Chen, Kang Liu, and Jun Zhao. 2018. Event detection via gated multilingual attention mechanism. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press.

Jian Liu, Yubo Chen, Kang Liu, and Jun Zhao. 2019. Neural cross-lingual event detection with minimal parallel resources. 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 738–748, Hong Kong, China. Association for Computational Linguistics.

Shulin Liu, Yubo Chen, Shizhu He, Kang Liu, and Jun Zhao. 2016. Leveraging FrameNet to improve automatic event detection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2134–2143, Berlin, Germany. Association for Computational Linguistics.

Shulin Liu, Yubo Chen, Kang Liu, and Jun Zhao. 2017b. Exploiting argument information to improve event detection via supervised attention mechanisms. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1789–1798, Vancouver, Canada. Association for Computational Linguistics.

Yaojie Lu, Hongyu Lin, Jin Xu, Xianpei Han, Jialong Tang, Annan Li, Le Sun, Meng Liao, and Shaoyi Chen. 2021. Text2Event: Controllable sequence-to-structure generation for end-to-end event extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2795–2806, Online. Association for Computational Linguistics.

Qing Lyu, Hongming Zhang, Elior Sulem, and Dan Roth. 2021. Zero-shot event extraction via transfer learning: Challenges and insights. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 322–332, Online. Association for Computational Linguistics.

Thien Huu Nguyen, Kyunghyun Cho, and Ralph Grishman. 2016. Joint event extraction via recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 300–309, San Diego, California. Association for Computational Linguistics.

Thien Huu Nguyen and Ralph Grishman. 2015. Event detection and domain adaptation with convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 365–371, Beijing, China. Association for Computational Linguistics.

Thien Huu Nguyen and Ralph Grishman. 2016. Modeling skip-grams for event detection with convolutional neural networks. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 886–891, Austin, Texas. Association for Computational Linguistics.

Trung Minh Nguyen and Thien Huu Nguyen. 2019. One for all: Neural joint modeling of entities and events. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Brendan O'Connor, Brandon M. Stewart, and Noah A. Smith. 2013. Learning to extract international relations from political context. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1094–1104, Sofia, Bulgaria. Association for Computational Linguistics.

Stephan Oepen, Marco Kuhlmann, Yusuke Miyao, Daniel Zeman, Dan Flickinger, Jan Hajič, Angelina Ivanova, and Yi Zhang. 2014. SemEval 2014 task 8: Broad-coverage semantic dependency parsing. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 63–72, Dublin, Ireland. Association for Computational Linguistics.

Tim O'Gorman, Kristin Wright-Bettner, and Martha Palmer. 2016. Richer event description: Integrating event coreference with temporal, causal and bridging annotation. In *Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016)*, pages 47–56, Austin, Texas. Association for Computational Linguistics.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Comput. Linguist.*, 31(1):71–106.

Vinodkumar Prabhakaran, Tomas By, Julia Hirschberg, Owen Rambow, Samira Shaikh, Tomek Strzalkowski, Jennifer Tracey, Michael Arrigo, Rupayan Basu, Micah Clark, Adam Dalton, Mona Diab, Louise Guthrie, Anna Prokofieva, Stephanie Strassel, Gregory Werner, Yorick Wilks, and Janyce Wiebe. 2015. A new dataset and evaluation for belief/factuality. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 82–91, Denver, Colorado. Association for Computational Linguistics.

James Pustejovsky, Patrick Hanks, Roser Saurí, Andrew See, Rob Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, and Marcia Lazo. 2003. The timebank corpus. *Proceedings of Corpus Linguistics*.

Rachel Rudinger, Aaron Steven White, and Benjamin Van Durme. 2018. Neural models of factuality. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 731–744, New Orleans, Louisiana. Association for Computational Linguistics.

Roser Saurí and James Pustejovsky. 2009. FactBank: a corpus annotated with event factuality. *Language resources and evaluation*, 43(3):227.

Philip A. Schrodt, Shannon G. Davis, and Judith L. Weddle. 1994. KEDS – a program for the machine coding of event data. *Social Science Computer Review*, 12(4):561 –587.

Lei Sha, Jing Liu, Chin-Yew Lin, Sujian Li, Baobao Chang, and Zhifang Sui. 2016. RBPB: Regularization-based pattern balancing method for event extraction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1224–1234, Berlin, Germany. Association for Computational Linguistics.

Zhiyi Song, Ann Bies, Stephanie Strassel, Tom Riese, Justin Mott, Joe Ellis, Jonathan Wright, Seth Kulick, Neville Ryant, and Xiaoyi Ma. 2015. From light to rich ERE: Annotation of entities, relations, and events. In *Proceedings of the The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 89–98, Denver, Colorado. Association for Computational Linguistics.

Gabriel Stanovsky, Judith Eckle-Kohler, Yevgeniy Puzikov, Ido Dagan, and Iryna Gurevych. 2017. Integrating deep linguistic features in factuality prediction over unified datasets. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 352–357, Vancouver, Canada. Association for Computational Linguistics.

Niklas Stoehr, Lucas Torroba Hennigen, Samin Ahbab, Robert West, and Ryan Cotterell. 2021. Classifying dyads for militarized conflict analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7775–7784, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sandesh Swamy, Alan Ritter, and Marie-Catherine de Marneffe. 2017. "i have a feeling trump will win..................": Forecasting winners and losers from user predictions on Twitter. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1583–1592, Copenhagen, Denmark. Association for Computational Linguistics.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, relation, and event extraction with contextualized span representations. 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 5784–5789, Hong Kong, China. Association for Computational Linguistics.

Xiaozhi Wang, Ziqi Wang, Xu Han, Zhiyuan Liu, Juanzi Li, Peng Li, Maosong Sun, Jie Zhou, and Xiang Ren. 2019. HMEAE: Hierarchical modular event argument extraction. 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 5777–5783, Hong Kong, China. Association for Computational Linguistics.

Ziqi Wang, Xiaozhi Wang, Xu Han, Yankai Lin, Lei Hou, Zhiyuan Liu, Peng Li, Juanzi Li, and Jie Zhou. 2021. CLEVE: Contrastive Pre-training for Event Extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6283–6297, Online. Association for Computational Linguistics.

Ralph M. Weischedel, Eduard H. Hovy, Mitchell P. Marcus, and Martha Palmer. 2017. Ontonotes : A large training corpus for enhanced processing.

Bishan Yang and Tom M. Mitchell. 2016. Joint extraction of events and entities within a document context. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 289–299, San Diego, California. Association for Computational Linguistics.

Sen Yang, Dawei Feng, Linbo Qiao, Zhigang Kan, and Dongsheng Li. 2019. Exploring pre-trained language models for event extraction and generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5284–5294, Florence, Italy. Association for Computational Linguistics.

Jiarui Yao, Haoling Qiu, Jin Zhao, Bonan Min, and Nianwen Xue. 2021. Factuality assessment as modal dependency parsing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1540–1550, Online. Association for Computational Linguistics.

Ying Zeng, Yansong Feng, Rong Ma, Zheng Wang, Rui Yan, Chongde Shi, and Dongyan Zhao. 2018. Scale up event extraction learning via automatic training data generation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press.

Junchi Zhang, Yanxia Qin, Yue Zhang, Mengchi Liu, and Donghong Ji. 2019a. Extracting entities and events as a single task using a transition-based neural model. In *International Joint Conference on Artificial Intelligence*.

Tongtao Zhang, Heng Ji, and Avirup Sil. 2019b. Joint Entity and Event Extraction with Generative Adversarial Imitation Learning. *Data Intelligence*, 1(2):99–120.

Yue Zhao, Xiaolong Jin, Yuanzhuo Wang, and Xueqi Cheng. 2018. Document embedding enhanced event detection with hierarchical and supervised attention. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 414–419, Melbourne, Australia. Association for Computational Linguistics.

Yang Zhou, Yubo Chen, Jun Zhao, Yin Wu, Jiexin Xu, and Jinlong Li. 2021. What the role is vs. what plays the role: Semi-supervised event argument extraction via dual question answering. In *AAAI Conference on Artificial Intelligence*.

A Appendix

A.1 Limitations

This work identifies specific issues and provides solutions to them. Recommendations 1 and 3 have solutions that could completely eliminate the issue that they address. The method that we introduce for recommendation 2 eliminates inconsistency in selecting the head of an argument extent; however, more ways of selecting the head may exist. Future work could explore additional ways of selecting the head in order to further reduce the chance that a correctly identified argument is considered as incorrectly identified.

A.2 Risks

The risks are the same as the risks for event extraction and information extraction. While a large literature, portions of which we reference, exists on ACE event extraction, less attention has been paid to its ethical and social implications. Sociopolitical events, which ACE often focuses on, may be of great interest to social scientists (e.g. the CASE workshop) as well as having government and military intelligence utility (presumably, an original motivation of the ACE program: while its original websites[11] and papers (Doddington et al., 2004) do not appear to explicitly specify a funding agency, they cite the earlier Message Understanding Conference (MUC) as its predecessor, whose proceedings explicitly cite DARPA as a sponsor (muc, 1991)). See, for example, Li et al. (2020b)'s ethical discussion of dual use issues for their partially ACE-based multimodal tracking/surveillance system.

A.3 Issues with the Current Literature for Identifying Arguments

In Section 4, we identified that several recent works since 2018, including some on zero-shot EE, do not

[11] https://www.ldc.upenn.edu/collaborations/past-projects/ace http://web.archive.org/web/20080303183132/https://www.nist.gov/speech/tests/ace/

evaluate the correctness of an argument by comparing it against all possible references to the argument within a sentence. We discuss more details about such works.

Wadden et al. (2019) state that "the ACE data set lacks coreference annotations," and the original released code[12] does not consider evaluating an argument against multiple references to the same argument. (As we note, ACE does in fact include significant coreference annotations.) Later, a third party added a software option to include clusters of entity spans, where a cluster contains spans of references referring to the same entity throughout a document, along with the event information. However, with this option, coreference resolution is still difficult because neither the entity information nor the event argument information in the pre-processed data includes an ID. While the pre-processed data includes entity and event argument spans, the spans may not completely match so mapping an event mention argument to an entity mention to check for multiple references using the pre-processed data becomes very difficult. Another third party also added code to gather coreference information corresponding to each event, but in the Github repository, one of Wadden et al. (2019)'s original authors states that both of these additions are unofficial.

We examine code bases of several works that design their pre-processing code similarly to Wadden et al. (2019) and find that they also do not collect all possible references to arguments from ACE (Du and Cardie, 2020; Lin et al., 2020; Lyu et al., 2021; Lu et al., 2021; Li et al., 2021). The Du and Cardie (2020) pre-processing code is most similar to the Wadden et al. (2019) pre-processing code, and the evaluation code does not compare arguments extracted by an EE method with ACE annotated references. Lin et al. (2020) and Lyu et al. (2021) state that they follow Wadden et al's pre-processing code and release their code bases. Although the code is more different than Du and Cardie (2020)'s code is, it does not gather multiple gold references for the same argument. Lyu et al. (2021) mention that some errors in the evaluation are attributable to this coreference issue. Further, Li et al. (2021) and Lu et al. (2021) both state that they follow Wadden et al. (2019)'s pre-processing and their respective code bases reflect this. Li et al. (2021) additionally state that they do not need to perform coreference resolution.

[12] https://github.com/dwadden/dygiepp

A.4 Exploration into the ACE Head and UD-based Head

We discuss the algorithm for identifying the UD-based head from the argument extent, and then show examples of the head that ACE identifies versus the head that the UD-based algorithm extracts.

A.4.1 Algorithm

The algorithm identifies the head of an argument extent in a way that is consistent with the Universal Dependency Parsing (UD) definition of head, but has slight modifications to suit the interpretation that a head could be an entire named entity and to work around possible well-known types of misparses by the UD formalism. The first step of the algorithm is to apply a tokenizer on the argument extent such that hyphens and apostrophes do not break words apart. Next, use SpaCy3 to construct a list of named entities that do not include the date, time, ordinal, or cardinal entity types. After, find the lowest common ancestor (LCA) for the argument extent. If the LCA is not within a named entity of the argument extent, select it as the head. Otherwise, select the named entity that the LCA is a substring of as the head.

The algorithm additionally handles two special cases that could complicate the UD selection of the appropriate head. If a null relativizer exists in an event argument, the UD parser may select a verb as the head. For example, in: "at least seven journalists killed covering the conflict", the parser selects "killed" as the head, which is incorrect. In addition, if a relative pronoun exists in an event argument, as in: "leader of the Iraq arms program who defected for a time", the UD parser may select the relativizer, "who", as the head. To work around these cases, the algorithm considers the argument extent to end after the first instance of a verb or relativizer pronoun that occurs after a noun (after a noun to avoid mis-identifying heads for cases such as: "these battered buildings").

We run the algorithm over all of the argument extents in ACE that are not of the form "[x] and/or [y]" since ACE has an exception of extracting two heads ([x] and [y]) from such extents, and find three mistakes out of a sample of 300. On the rare single-word case that a mistake occurs, the argument span usually contains a noun compound with spaces (most such noun compounds do not indicate a mistake), and none of these spans contain null relativizers.

A.4.2 Contradictions

We show surprising discrepancies between the head that ACE identifies and the head that the UD-based algorithm identifies with respect to an argument extent below. Similar to the examples in Figure 3 of the main paper, the head that ACE identifies is in red brackets and the head that the UD-based algorithm identifies is in blue brackets.

the [Houston [Center]$_{ACE}$]$_{AUT}$
[Wall [street]$_{AUT}$]$_{ACE}$
[aol time [warnerings]$_{AUT}$]$_{ACE}$
[f-14 [aircraft]$_{ACE}$]$_{AUT}$
another [half-[brother]$_{ACE}$]$_{AUT}$ of saddam hussein
[neither]$_{AUT}$ of the [women]$_{ACE}$
the [[Office]$_{ACE}$ of the President]$_{AUT}$
the [[president]$_{AUT}$]$_{ACE}$-elect of the American Medical Association
several [[parts]$_{ACE}$]$_{AUT}$ of southern Iraq
[hundreds]$_{AUT}$ of [civilians]$_{ACE}$ in East Timor
a [[warren]$_{ACE}$]$_{AUT}$ of cells
[thousands]$_{AUT}$ of U.S. [troops]$_{ACE}$
the [[Shah]$_{ACE}$ of Iran]$_{AUT}$
the [U.S. Army [7th Cavalry]$_{ACE}$]$_{AUT}$
[American [Marines]$_{ACE}$]$_{AUT}$
two [U.S. [Marines]$_{ACE}$]$_{AUT}$ killed in combat
21-year- old [Marine Corporal [Randall Kent Rosacker]$_{ACE}$]$_{AUT}$
[delma [banks]$_{AUT}$]$_{ACE}$
the [national youth and student peace [coalition]$_{AUT}$]$_{ACE}$
[persian [gulf]$_{AUT}$]$_{ACE}$
the [center]$_{ACE}$ of the second largest city in iraq, [basra]$_{AUT}$
the [urbuinano [island]$_{AUT}$]$_{ACE}$
the [catholic [church]$_{ACE}$]$_{AUT}$ in phoenix, arizona
two very strong – [militant groups]$_{ACE}$
British [Desert [Rats]$_{AUT}$]$_{ACE}$
the [Alfred P. Murrah federal [building]$_{AUT}$]$_{ACE}$
his [ex-[wife]$_{ACE}$]$_{AUT}$
[tight [ends]$_{AUT}$]$_{ACE}$
[9]$_{AUT}$ [more]$_{ACE}$
[19]$_{AUT}$ [more]$_{ACE}$
[second-[graders]$_{ACE}$]$_{AUT}$

A.5 ACE Experiment Details

To extract statistics about coreference, we modify Wadden et al's pre-processing code. In the analysis, we omit one document due to preprocessing issues and do not consider times and values as arguments; only entities, which is consistent with most of the literature that we reviewed.

From the results in Table 2, we observe that the selected event mention argument does seem to follow a specific pattern; it does not seem to prefer being a named entity, nor consistently be the first of the references to appear in a sentence; etc.

If multiple non-duplicate refs exist in the same sentence, the percent that:	**Excl. Pron.**	**Incl. Pron.**
the event arg is a named entity, given $\geq$ 1 reference is a named entity	67.63	60.73
the event arg is not a named entity, given $\geq$ 1 reference is a named entity	32.37	39.27
the event arg is the first of those references in the sentence	47.90	56.32
the event arg is not the first of those references in the sentence	52.10	43.68
the event arg is not a relativizer pronoun, given $\geq$ 1 reference is a relativizer pronoun	n/a	80.63
the event arg is a relativizer pronoun, given $\geq$ 1 reference is a relativizer pronoun	n/a	19.37
the event arg is not a different pronoun, given $\geq$ 1 reference is a different pronoun	n/a	67.46
the event arg is a different pronoun, given $\geq$ 1 reference is a different pronoun	n/a	32.54

Table 2: Percentage information about the event mention argument in the case that multiple non-duplicate references ($\geq$ 2) to the same entity exist *in the same sentence*. A relativizer pronoun includes "who", "which"; etc while a different pronoun includes "he", "her"; etc. We extract this number in cases where arguments can be pronouns and where they cannot be.

A.6 Literature Review Details

To aim toward fair comparison among EE methods, works use ACE to evaluate them in three general ways. Only the earliest papers (Ji and Grishman, 2008; Liao and Grishman, 2010; Hong et al., 2011) use the first split **(A)**, where the evaluation uses all of the text data and 33 separate event subclasses, ignoring the event classes, and where the test set contains 40 newswire texts, the development set contains 10 newswire texts, and the rest of the texts belong to the training set. The second split (**B**) is an improvement upon the first, with the only difference of using *30 randomly selected* texts in the development set. A zero-shot evaluation of this split variety ignores the training set. A third split variety (**C**) is for a specific application of event extraction which focuses more on the generalization ability across different domains; in this split, the source domain is news, half of bc is the development set, and the remaining data makes up the test set. Three papers that we reviewed use split **(A)** (Ji and Grishman, 2008; Liao and Grishman, 2010; Hong et al., 2011), at least 28 papers use split **(B)** (Li et al., 2013; Nguyen and Grishman, 2015; Chen et al., 2015; Nguyen et al., 2016; Yang and Mitchell, 2016; Nguyen and Grishman, 2016; Feng et al., 2016; Liu et al., 2016; Huang et al., 2016; Sha et al., 2016; Chen et al., 2017; Liu et al., 2017b; Zhao et al., 2018; Liu et al., 2018, 2019; Zhang et al., 2019b; Wang et al., 2019; Zhang et al., 2019a; Yang et al., 2019; Nguyen and Nguyen, 2019; Wadden et al., 2019; Liu et al., 2020; Li et al., 2020a; Ahmad et al., 2021; Lu et al., 2021; Lyu et al., 2021; Wang et al., 2021; Zhou et al., 2021), some for few-shot or zero-shot evaluations use a different, contrived split (e.g. Huang et al. (2018)) and others use both split (B) and a different split (e.g. Du and Cardie (2020)).

In addition, most works use the evaluation criteria that 1. The *event trigger is considered correct* when its offsets match a gold trigger and event class is correct and 2. An *argument is considered correct* when its offsets and event class match a gold argument and its event role is correct. However, the criteria does not include many more details and is not in formal math notation, allowing discrepancies in the way that different works implement them.