EventOA: An Event Ontology Alignment Benchmark Based on FrameNet and Wikidata

Shaoru Guo[1], **Chenhao Wang**[1,2], **Yubo Chen**[1,2*], **Kang Liu**[1,2,3], **Ru Li**[4] and **Jun Zhao**[1,2]
[1]The Laboratory of Cognition and Decision Intelligence for Complex Systems,
Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3]Beijing Academy of Artificial Intelligence, Beijing, China
[4]School of Computer and Information Technology, Shanxi University, Taiyuan, China
{shaoru.guo,chenhao.wang,yubo.chen,kliu,jzhao}@nlpr.ia.ac.cn, liru@sxu.edu.cn

Abstract

Event ontology provides a shared and formal specification about what happens in the real world and can benefit many natural language understanding tasks. However, the independent development of event ontologies often results in heterogeneous representations that raise the need for establishing alignments between semantically related events. There exists a series of works about ontology alignment (OA), but they only focus on the entity-based OA, and neglect the event-based OA. To fill the gap, we construct an *<u>E</u>vent <u>O</u>ntology <u>A</u>lignment* (**EventOA**) dataset based on FrameNet and Wikidata, which consists of 900+ event type alignments and 8,000+ event argument alignments. Furthermore, we propose a multi-view event ontology alignment (MEOA) method, which utilizes description information (i.e., name, alias and definition) and neighbor information (i.e., subclass and superclass) to obtain richer representation of the event ontologies. Extensive experiments show that our MEOA outperforms the existing entity-based OA methods and can serve as a strong baseline for EventOA research.

1 Introduction

Event ontology is crucial for understanding human behavior and has become a new paradigm for describing knowledge in the Semantic Web by providing a shared and formal specification about what happens in the real world (Brown et al., 2017). As shown in Figure 1, event `Attack` accurately describes the action in which someone attempts to injure another organism with many-sided arguments such as "*assailant*", "*victim*", "*weapon*", and so on. It has been recognized as useful for tasks like information extraction (Wimalasuriya and Dou, 2010), web service (Li and Yang, 2008) and automatic question answering (Lopez et al., 2011). Thus a remarkable number of event ontologies have been created such as FrameNet (Baker et al., 1998), VerbNet (Kipper et al., 2007), Wikidata (Erxleben et al., 2014) and ACE (Doddington et al., 2004). However, the independent development of event ontologies often results in heterogeneous representations that hinder the knowledge integration.

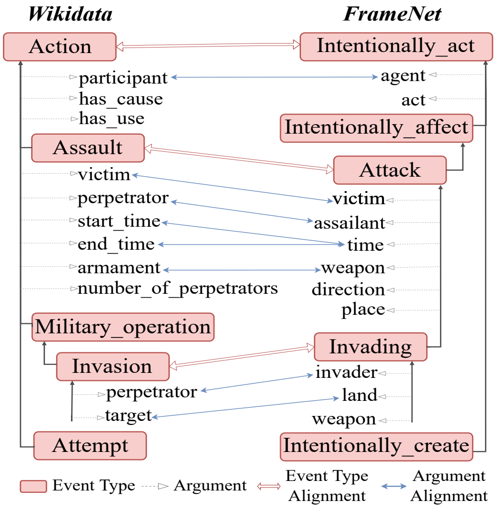


Figure 1: Examples of FrameNet and Wikidata ontology alignment.

Driven by the ontology alignment evaluation initiative (OAEI) [1] (Pour et al., 2022), many datasets (Bodenreider et al., 2005; Svátek and Berka, 2005; Karam et al., 2020) and methods (Jiménez-Ruiz et al., 2013; Faria et al., 2013; Iyer et al., 2021) have been proposed for ontology alignment (OA). However, almost all datasets and methods so far focus on entity ontologies, which are known for sharing knowledge about entities such as people, organizations and products (Bodenreider et al., 2005; Zamazal and Svátek, 2017). In contrast, event ontologies, which provide nexus for related entities/arguments with a higher semantic granularity, are more useful for language understanding tasks (Brown et al., 2017), but there is little attempt to

[*] Corresponding Author.

[1] A public platform to collect datasets for ontology alignment tools, and a regular evaluation of those tools since 2004.

Findings of the Association for Computational Linguistics: ACL 2023, pages 10038–10052
July 9-14, 2023 ©2023 Association for Computational Linguistics

tackle the problem of event-based OA.

To address the above issues, we take FrameNet (Baker et al., 1998) and Wikidata (Erxleben et al., 2014) as examples to explore the alignment between event ontologies. As illustrated in Figure 1, for Wikidata and FrameNet, OA systems need to establish correspondences between event types such as `Assault` vs. `Attack`, and correspondences between event arguments such as "*armament*" vs. "*weapon*". We choose FrameNet and Wikidata as the data sources for the following reasons.

First, establishing correspondences between FrameNet and Wikidata is meaningful as it will help to obtain an integrated event ontology with **high coverage and quality**. On one hand, Wikidata is a widely used world knowledge base contributed by the community, which has a large number of events but with a confusing hierarchy (Pellissier Tanon et al., 2020). On the other hand, FrameNet is an excellent repository of linguistic knowledge designed by linguists, which has a logically clean hierarchy but with limited events. Specifically, Wikidata contains 290K events that cover a wide range of domains, including disaster, sport, election, etc. However, the hierarchy in Wikidata is confusing as anyone can edit relations between events. For example, `Writing` is a subclass of `Artistic_creation`, while `Carving` is a subclass of `Change`. So a query for `Artistic_creation` would find the `Writing` but not `Carving`. In fact, both `Writing` and `Carving` are `Artistic_creation` activities. In contrast, FrameNet has an agreed-upon hierarchy that cannot be changed unless by the agreement of linguists, but FrameNet does not cover latest major events such as `2022_FIFA_World_Cup`. Thus an ontology that reconciles the rigorous hierarchy of FrameNet with the rich events of Wikidata is valuable for applications in the Semantic Web.

Second, establishing correspondences between FrameNet and Wikidata is challenging due to the **semantic diversity** of lexemes described the event type and argument (i.e., *polysemy* and *synonymy*). (i) *Polysemy*, which refers to the phenomenon that ontologies use the same lexeme to describe events with different purposes, e.g., `Motion` in FrameNet describes the everyday events of "*Agents change in position over time*", while `Motion_Q452237` in Wikidata describes "*parliamentary motion*" that happens throughout mankind history. Polysemy also occurs when the semantics of arguments vary from event type to event type (Li et al., 2006). As shown in Figure 1, argument "*perpetrator*" of event `Assault` corresponds to "*assailant*" of event `Attack`, while argument "*perpetrator*" of event `Invasion` corresponds to "*invader*" of event `Invading`. How to identify different semantics of the same lexeme is a challenging issue. (ii) *Synonymy*, which refers to the phenomenon that FrameNet and Wikidata use different lexemes to refer the same event types or arguments. As shown in Figure 1, `Assault` and `Attack` express the same event type with different lexemes, meanwhile "*armament*" and "*weapon*" express the same event argument with different lexemes. Thus it is critical to build complex correspondences that are semantics related but with different lexemes.

To this end, in this paper, we build an event ontology alignment dataset based on FrameNet and Wikidata. This dataset is named as EventOA and composed of two sub-datasets: *event type alignment* and *event argument alignment*. We extensively evaluate existing OA methods, but they are far from solving EventOA. Thus we propose a multi-view event ontology alignment (MEOA) method by utilizing multi-view information of event ontologies, which we believe would serve as a strong baseline for EventOA. We further propose a reasonable evaluation metrics for EventOA with type alignment and argument alignment. Our contributions are as follows:

- We construct EventOA, a real world event ontology alignment dataset based on FrameNet and Wikidata, which consists of two subtasks, namely, event type alignment and event argument alignment. In addition, we devise evaluation metrics for the two subtasks to assess alignment quality.

- We propose a multi-view event ontology alignment (MEOA) method, which utilizes multi-view information to model the representation of event ontology and thus can better resolve the semantic diversity problem.

- We conduct extensive evaluations of existing entity-based OA methods and our MEOA method. Experiment results show that our MEOA method outperforms the entity-based methods and achieves the SOTA performance, which can serve as a strong baseline for EventOA research. We also conduct a detailed error analysis to provide insights to future work.

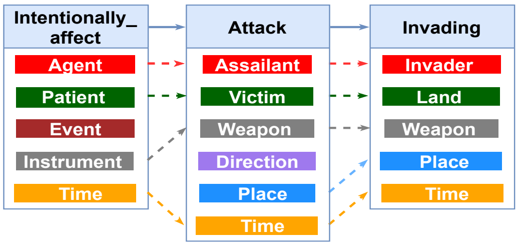


Figure 2: An example of F-to-F and FE-to-FE. Solid lines represent F-to-F. Dash lines represent FE-to-FE.

	FrameNet	Wikidata
Type	1,221	12,159
Argument	11,428	257,498
Avg. Argument	9.4	21.2
Maximum Depth	10	12
Type Alignment	905	
Argument Alignment	8,650	

Table 1: Statistics of the EventOA dataset.

2 Data Construction of EventOA

We construct our dataset in four stages: FrameNet-based ontology collection, Wikidata-based ontology collection, automatic alignment candidate selection and human annotation.

2.1 FrameNet-based Ontology Collection

FrameNet (Fillmore, 1976; Baker et al., 1998) is a linguistic resource constructed by linguists, which describes everyday events with agreed-upon *inheritance relations*. Thus we construct FrameNet-based ontology by collecting event types and arguments from FrameNet and building the hierarchy based on the inheritance relations.

In particular, Frame (Guan et al., 2021) is defined as schematic representation of a situation. Frame Elements (FEs) are frame-specific defined semantic roles. Lexical Units (LUs) are set of words grouped by their senses, and belong to a particular frame. Frames are linked by frame-to-frame relations (F-to-F) such as "Inheritance" and "Subframe". And the relations between FEs are the same as the corresponding relations between frames. For instance, in Figure 2, FE "*assailant*" inherits from FE "*agent*" as frame `Attack` inherits from frame `Intentionally_affect`.

Based on the FrameNet inheritance relations, we construct FrameNet-based ontology, where the *frame* can be viewed as *event type*, the *FE* can be viewed as *event argument*, and F-to-F and FE-to-FE respectively reflect the relations among events and arguments. We build the RDFS schema for FrameNet according to FrameBase, which translates frame, frame element, F-to-F and FE-to-FE into RDFS counterparts (Rouces et al., 2015).

2.2 Wikidata-based Ontology Collection

Wikidata (Erxleben et al., 2014) is a community effort where anybody can contribute facts, resulting in a confusing knowledge base including *circle paths*, *useless events* and *incomplete arguments*. Thus we construct Wikidata-based ontology by processing the above confusions as follows:

(1) *Data acquisition*. Inspired by Gottschalk and Demidova (2019), we run the SPARQL query in Figure 4 of Appendix B to select subclasses of Wikidata's "occurrence" as our event dataset. (2) *Circle-path filtration*. For an event in the circle, we only retain the path with the smallest depth to the root "occurrence". (3) *Useless-events deletion*. For each path, we discard the classes that have less than 10 direct instances and at the same time directly assert their children as subclasses of their parents for keeping the hierarchy. (4) *Arguments completion*. Given an event, we collect all its direct instances and use the union set of instance's properties as its arguments, as shown in Figure 1, arguments of event `Assault` (e.g., "*victim*" and "*perpetrator*") are obtained from its instances.

2.3 Automatic Alignment Candidate Selection

Given the two event ontologies (FrameNet and Wikidata), our goal is to identify correspondences between event type (frame and class) and event argument (FE and property) [2]. To facilitate efficiency of annotation, we adopt some heuristic and automatic methods to select alignment candidates.

Event type candidate selection aims to select candidate frames in FrameNet for a given event class in Wikidata [3]. We apply *Frame-based* and *LU-based* methods combined with *Similarity-based* method for event type candidate selection.

Frame-based method selects frames that are same as any forms of the event class in Wikidata as its candidates.

[2] In order to better distinguish the data sources, we'll use specific concepts to represent **event type** (i.e., *frame* of FrameNet and *class* of Wikidata) and **argument** (i.e., *frame element* of FrameNet and *property* of Wikidata), respectively.

[3] It does no matter which resource is fixed, we fix Wikidata and select candidates from FrameNet in this work as the number of frames is smaller than classes in Wikidata.

Datasets			Type/Class	Arg/Pro	Type/Class Alignment	Arg/Pro Alignment	Total Alignment
Track		Ontology					
Entity OA	Conference	Edas	104	50	12	3	15
		Sigkdd	49	28			
	Anatomy	AMA	2,744	2	1,544	0	1,544
		NCI-A	3,304	3			
	Biodiv	ENVO	6,566	136	822	0	822
		SWEET	4,533	0			
EventOA (Ours)		FrameNet	1,221	11,428	905	8,650	9,555
		Wikidata	12,159	257,489			

Table 2: Comparison between entity and event OA datasets. Arg and Pro refer to argument and property, respectively. Note event includes type and argument, and entity includes class and property.

LU-based method selects frames whose lexical units are same as any forms of the event class in Wikidata as its candidates.

Similarity-based method is used to amend the candidates number when the total number of candidates selected by Frame- and LU-based methods is less than 15. It selects candidates by computing similarity between class representation S_c and frame representation S_f. We use frame name F_n and lexical unit F_{lu} to build the representation S_f (Guo et al., 2020). F_{lu} representation is obtained by averaging the embedding of all LUs lu in a frame, i.e., $F_{lu} = \frac{1}{M}\sum_{i=1}^{M} lu_i$. M is the total number of LUs of the frame. S_c, F_n and lu are the pre-trained Glove (Pennington et al., 2014).

Event argument candidate selection attempts to construct candidates for each property in Wikidata with FEs. We apply a *Relation-aware Attention Mechanism* for argument candidate selection.

For each property in Wikidata, we use the FEs under the corresponding frame as candidates and rank FEs by calculating similarity between property p and FE fe. Specifically, we integrate the nominal and relational perspectives of a FE for a more comprehensive representation as shown in Equation (1). FE_n represents the nominal perspective, and $\sum_{w=1}^{W} att(FE_w) \cdot FE_w$ represents the relational perspective. We utilize FE-to-FE relations to model FEs relational perspective with attention schema. Given a FE, $FE^+ = \{FE_1, FE_2, \ldots, FE_W\}$ represents its expanded FEs, including all FEs that can be linked to FE through FE-to-FE relations. Note attention schemes have been designed to emphasize relevant FEs, avoiding the influence from less relevant but linked FEs.

$$fe = FE_n + \sum_{w=1}^{W} att(FE_w) \cdot FE_w \qquad (1)$$

$$att(FE_w) = \frac{exp(FE_n \cdot FE_w)}{\sum_{k=1}^{W} exp(FE_n \cdot FE_k)} \qquad (2)$$

where FE_n is FE name representation, and W stands for the total number of FEs in FE^+. We utilize the same method to obtain the p.

2.4 Human Annotation

We obtain candidates through above process for an event, but the semantic distinctions among candidates are subtle, so it is difficult to automatically select the best alignment to ensure the quality.

To create a gold-standard dataset of event ontology alignment, three graduate students who are familiar with Wikidata and FrameNet are invited to label the class with appropriate frame and label property with appropriate FE using our internal annotation platform, a screenshot of annotation interface is shown in Appendix E. They work independently and we adopt the majority vote for deciding the final correspondences (if disagreement appears). The mean inter-annotation agreement computed by Cohen's Kappa is 82.4%, indicating a high annotation quality. Examples of alignments are provided in Appendix A.

2.5 Data Analysis

Statistics of EventOA. As shown in Table 1, the FrameNet-based ontology contains 1,221 event types and 11,428 arguments, and the Wikidata-based ontology includes 12,159 event types and 257,498 arguments. By automated selection and human annotation, EventOA dataset contains 905 event type alignments and 8,650 event argument alignments, which is rich enough to promote the research of event ontology alignment.

Comparison between entity and event OA datasets. We compare EventOA with existing widely-used EntityOA datasets in Table 2. From the table, we can observe that: (1) *The size of FrameNet ontology and Wikidata ontology is significantly different*. Concretely, the size of entity ontologies in each track have similar magnitudes (e.g.,

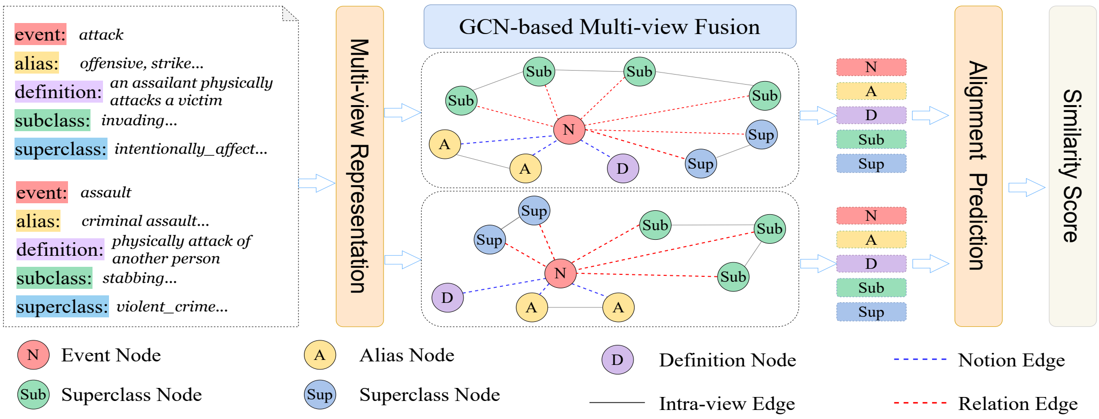

event: attack
alias: offensive, strike...
definition: an assailant physically attacks a victim
subclass: invading...
superclass: intentionally_affect...
event: assault
alias: criminal assault...
definition: physically attack of another person
subclass: stabbing...
superclass: violent_crime...
Multi-view Representation
GCN-based Multi-view Fusion
Sub
Sup
N
A
D
Alignment Prediction
Similarity Score
Event Node
Alias Node
Definition Node
Notion Edge
Superclass Node
Superclass Node
Intra-view Edge
Relation Edge

Figure 3: Illustration of Multi-view Event Ontology Alignment (MEOA) Method.

2,744 (AMA) vs. 3,304 (NCI-A)), while the size of Wikidata is larger than that of FrameNet (i.e., 1,211 vs. 12,159). This is because the construction methods of FrameNet and Wikidata are different (experts vs. community), which leads to diverse representation for events. (2) *EventOA has rich arguments*. Concretely, EntityOA datasets contain very little properties (e.g., 136 in ENVO), while our EventOA has a larger number of arguments (e.g., 257,489 in Wikidata). The reason is that arguments are defined specifically to each event type and thus lead to the diversity representation.

3 Multi-view Event Ontology Alignment (MEOA)

To solve the semantic diversity problems, we design MEOA, which establishes correspondences between event ontologies by utilizing multi-view information, as shown in Figure 3.

3.1 Multi-view Representation (MR)

MR aims to represent FrameNet and Wikidata from five different views, including name, alias, definition, subclass and superclass [4]. We choose these views as they can well describe the *description* (name, alias and definition) and *neighbor* (subclass and superclass) information for an event.

Denote the different meaningful views as $P = \{P_1, P_2, \ldots, P_i, \ldots, P_N\}$, and N is the number of views. $P_i = \{P_{i1}, P_{i2}, \ldots, P_{ij}, \ldots, P_{iM}\}$, P_{ij} is the j-th element of view P_i. For P_{ij}, we feed its information $P_{ij} = \{w_{ij1}, w_{ij2}, \ldots, w_{ijk}, \ldots, w_{ijK}\}$ into the transformer-based encoder (Vaswani et al., 2017) to generate p_{ij}.

$$p_{ij} = \sum_{k=1}^{K} embedding(w_{ijk}) \qquad (3)$$

where w_{ijk} represents the k-th word in the P_{ij} , and K is the total number of words.

3.2 Multi-view Fusion (MF)

MF aims to integrate multi-view embeddings to get a more meaningful representation.

Intuitively, the combination of multi-view embeddings can strengthen the event representation. To model the multi-view information and interactions among different views, a multi-view Event Ontology Graph (EOG) is constructed.

EOG has five different kinds of nodes that correspond to the five views described in Section 3.1. There are two types of edges in EOG:

Intra-view Edge: Nodes referring to the same view are connected with intra-view edges. In this way, the interaction among different nodes of the same view could be modeled.

Inter-view Edge: Different views are connected with inter-view edges if they belong to the same event, which can be further divided into *Description edge* and *Neighbor edge*.

Description Edge connects alias view and definition view to the name view. The rationale is that equivalent events tend to share similar or even the same notions.

Neighbor Edge connects superclass view and subclass view to the name view. The rationale is that equivalent events tend to be neighbored by equivalent events.

We apply Graph Convolution Network (GCN) (Kipf and Welling, 2017) on EOG to aggregate information. Formally, the hidden representation

[4] We model argument name representation with the help of its event type by directly summing up their name embeddings.

for each node at *(l+1)*th layer is computed by:

$$H^{l+1} = \phi(\widetilde{D}^{-\frac{1}{2}}\widetilde{A}\widetilde{D}^{-\frac{1}{2}}H^{l}W^{l}) \tag{4}$$

where $\widetilde{A} = A + I$ is the adjacency matrix of the graph EOG with added self-connections, I is the identity matrix, $\widetilde{D}$ is the diagonal node degree matrix of $\widetilde{A}$, $\phi(\cdot)$ is ReLU function, and W^l denotes learn-able parameters in *l*-th layer.

3.3 Alignment Prediction (AP)

AP aims to establish correspondences between semantically related events from different ontologies.

For FrameNet (E_1) and Wikidata (E_2), we define the correspondence between two events $e_1 \in E_1$ and $e_2 \in E_2$ as the three-element tuple, i.e., $T =< e_1, e_2, s >$, where $(f, fe) \in e_1$, $(c, p) \in e_2$, and $s \in [0, 1]$ is a score indicating the degree to which e_1 and e_2 are equivalent. Event type (f,e) and argument (fe,p) representation are obtained from MF module. We respectively compute the confidence score of type alignment S_t and argument alignment S_a, and use mean squared error as the loss to train our model inspired by Iyer et al. (2021). We take event type alignment as an example to elaborate the process.

$$S_t(f_i, c_j) = cos\_sim(f_i, c_j) \tag{5}$$

$$\mathcal{L}_t = \frac{1}{T}\sum(S_t(f_i, c_j) - G_t(f_i, c_j))^2 \tag{6}$$

where $S_t(\cdot)$ denotes the confidence score of event type alignment, and $G_t(\cdot)$ denotes the ground truth label which is 1 if $f_i \equiv c_j$ and 0 otherwise. Note the process of argument alignment prediction is same as the process of type alignment, and the details can be found in Appendix C.

4 Experiments

This section provides experiment details, i.e., evaluation metrics, baselines, results, and their analysis.

4.1 Evaluation Metrics

Inspired by Faria et al. (2013) and Ji and Grishman (2008), we define two standards to determine the *correctness* of alignment (type and argument):

- *An event type alignment is correctly identified* if it matches a reference event type alignment.
- *An event argument alignment is correctly identified* if the event type alignment and argument alignment match any of the reference argument alignments.

Following prior work (Faria et al., 2013; Iyer et al., 2021), we use Precision (P), Recall (R) and F-measure (F1) to evaluate the performance.

$$P = \frac{M_{Out} \cap M_{RA}}{M_{Out}} \tag{7}$$

$$R = \frac{M_{Out} \cap M_{RA}}{M_{RA}} \tag{8}$$

$$F1 = 2\frac{P \cdot R}{P + R} \tag{9}$$

where M_{Out} are the system's output alignments and M_{RA} are reference (a.k.a. gold) alignments.

4.2 Data Splitting and Baseline Models

We consider two settings for data splitting: (i) the entire data is treated as test set, which is suitable for comparing unsupervised OA methods; (ii) the entire data is split into training, validation and test sets in 70:10:20 ratio, which can be used for evaluating supervised OA methods.

We compare MEOA with various baselines: (i) unsupervised OA methods, namely AML (Faria et al., 2013), LogMap (Jiménez-Ruiz and Cuenca Grau, 2011; Jiménez-Ruiz et al., 2020) and Wiktionary (Portisch et al., 2019); (ii) supervised OA methods such as Word2Vec + classifier (He et al., 2022) and VeeAlign (Iyer et al., 2021). We choose these baselines based on their *top performing* and *open-source availability*. Note Word2Vec + classifier method concatenates embeddings of two event types or arguments and feeds them to a classifier trained on a training set to output an alignment score. For our unsupervised method MEOA-sum, we directly perform a sum operation rather than the GCN to unify different views into a single vector for each event ontology to calculate the alignment score.

Details about baselines and the implementation of our MEOA are provided in Appendix D.

4.3 Results and Discussion

We demonstrate the effectiveness of the proposed MEOA method and the challenges of EventOA.

Performance comparison of different methods on EventOA. Table 3 presents the performance of our MEOA model on EventOA (including event type and event argument alignment) compared with top performer unsupervised/supervised entity-based OA methods. From the table, we can see that: (1) Our MEOA method achieves the highest F-measure on EventOA and significantly outperforms the baselines for t-test (p-value<0.05),

Method	Event Type Alignment			Event Argument Alignment		
	Precision	Recall	F-measure	Precision	Recall	F-measure
Unsupervised OA Methods(%)						
LogMap	27.89	25.94	26.87	34.71	7.27	12.02
AML	33.52	19.07	24.30	34.87	13.59	19.56
Wiktionary	23.73	53.77	32.92	33.64	11.72	17.38
MEOA-sum(Ours)	41.76	40.68	**41.21**$^{\uparrow(8.29)}$	40.23	26.79	**32.16**$^{\uparrow(12.60)}$
Supervised OA Methods(%)						
Word2Vec + classifier	42.03	35.58	38.53	29.02	21.50	24.70
VeeAlign	47.91	72.48	57.69	40.25	37.56	38.86
MEOA(Ours)	58.22	75.81	**65.86**$^{\uparrow(8.17)}$	38.92	54.89	**45.55**$^{\uparrow(6.69)}$

Table 3: Results on EventOA. Comparison of our proposed MEOA method with top performing unsupervised/supervised entity-based OA methods on event type and event argument alignment.

Dataset	Precision	Recall	F-measure
Conference	68.48	67.91	68.19
Biodiv	84.48	82.56	83.50
Anatomy	89.62	85.73	87.63
EventOA (Ours)	58.22	75.81	65.86

Table 4: Experimental results of MEOA on three entity-based datasets and our EventOA, indicating the challenge of EventOA and generalization of MEOA.

Method	Precision	Recall	F-measure
MEOA	58.22	75.81	65.86
w/o definition	57.36	69.51	62.85
w/o subclass	57.72	68.13	62.49
w/o superclass	58.74	63.88	61.20
w/o alias	60.64	59.30	59.96

Table 5: Ablation study on event type alignment.

suggesting that our MEOA can utilize multi-view information to model event ontology that is better suited for EventOA. (2) MEOA outperforms the MEOA-sum, which indicates that using GCN for multi-view fusion is, unsurprisingly, more effective than direct summation. (3) The performances of entity-based OA methods are lower and not satisfying on EventOA, which indicates that our EventOA dataset is challenging for the existing OA systems. The reason may be that entity-based OA methods mainly consider surface features such as overlapped sub-strings, and fail to capture the semantics information. However, the string features of events in EventOA such as `Engagement` and `Hostile_encounter` are not similar, but their semantics are related.

Performance of MEOA on different OA datasets. We test our MEOA on three entity-based datasets in Table 4. From the table, we can observe that: (1) our MEOA achieves better performance on entity-based OA datasets, which manifests the *high generalization* ability of our MEOA method. (2) Although MEOA achieves better performance on EventOA, experimental results still demonstrate a gap between event-based OA (65.86% F-measure) and entity-based OA (e.g., 87.63% F-measure on Anatomy), indicating our EventOA is more *challenging* and more research efforts on event-based OA are needed in future work [5].

4.4 Ablation Study

We conduct ablation studies on event type alignment in Table 5. From the table, we can see that: (1) Eliminating any of the semantic views, namely alias, definition, superclass and subclass, would hurt the performance, which validates the effectiveness of several views in our model.(2) Comparing the four models with our MEOA model, we can observe that w/o *alias* hurts the performance most. This intuitively makes sense, since *aliases* are description representation that can directly reflect the semantics of an event ontology. (3) After ablating superclass and subclass, the F-measure drops by 4.66% and 3.37%, which demonstrates that the neighbor information is valuable for capturing the semantic correlation between event ontologies.

[5] The results of entity-based systems on the three entity-based datasets (i.e., Conference, Biodiv and Anatomy) can be obtained from OAEI 2022 campaign https://oaei.ontologymatching.org/2022/.

	Wikidata	FrameNet
Type	War	Hostile_encounter
	Human_migration	Quitting_a_place
	Wedding	Forming_relationships
Arg	Armament	Weapon
	Inception	Time_of_creation

Table 6: Examples of alignments that are correctly predicted by MEOA. Arg refers to argument.

4.5 Case Study

To show the effects of our MEOA model, Table 6 shows some cases of alignments that are correctly predicted by our MEOA but not by AML or LogMap. We can clearly see that, our MEOA method can resolve semantic diversity problems by capturing the implicit connection between ontologies. For instance, MEOA knows `War` corresponds to `Hostile_encounter` according to the *alias* information (i.e., `War` matches an alias in `Hostile_encounter`), as well as "*Armament*" and "*Weapon*" by utilizing arguments' definition and relation information. This demonstrates the strength of MEOA for modeling multi-view information to improve the performance of the alignment.

4.6 Error Analysis

Table 7 shows examples of error cases. Note that these cases are also incorrectly predicted by LogMap and AML. (1) **Ambiguity**, where distinctions between events can be relatively subtle. For instance, `Medical_intervention` differs from `Cure` mainly in the effect of the treatment, i.e., `Medical_intervention` deals only with attempts to *alleviate* a `Medical_condition`, whereas *Cure* deals with situations in which the `Medical_condition` has been *cured*. So it is difficult for a model to distinguish which event corresponds to event `Treatment`. (2) **Compound Word**, where event types are formed with two or more words that make it difficult to derive accurate representations for them. For example, `Deliberate_murder` refers to `Killing` as "*deliberate*" is used to modify "*murder*". However, words in `Surgical_operation` are used together to take on a new meaning that refers to `Medical_intervention`. (3) **Spurious Correlation**, where relations of arguments are too fraudulent for models to see through their spurious relationships and consequently resulting in poor generalization, e.g., "*Vehicle*" relates to "*Speed*" in many cases, so models learn this spurious correlation and cannot generalize to "*Impact*" where "*Vehicle*" refers to "*Impactors*". (4) **Same Category**, where models fail to discriminate semantics among arguments whose categories are same. For "*Director*", model outputs "*Performer*" when the gold argument is "*Personnel*" as both of them belong to *people* category, and models cannot further discriminate semantics between them.

5 Related Work

As this work involves datasets and methods about ontology alignment, we review key related works in these areas.

The OAEI has been the foremost venue for researchers focused on OA task, so we begin with a survey of datasets have been used in the OAEI. Anatomy is one of the longest running tracks in the OAEI, which consists of human and mouse anatomy ontologies from the biomedical domain and have been manually matched by medical experts (Bodenreider et al., 2005; Dragisic et al., 2017). Biodiv is particularly useful for biodiversity and ecology research (Karam et al., 2020). Conference is a collection of ontologies from the same domain of organizing conferences using complex definitions (Svátek and Berka, 2005). All of these datasets are about entity-based OA, but neglect the event-based OA.

Methods of OA can be classified into feature-based methods and deep learning based methods. Feature-based methods are typically based on lexical matching. Among these systems, AgreementMakerLight (AML) (Faria et al., 2013) and LogMap (Jiménez-Ruiz and Cuenca Grau, 2011) are two classic and leading systems in many OAEI tracks and other tasks (Kolyvakis et al., 2018). Wiktionary (Portisch et al., 2019) is another top performing system for multilingual OA. Recently, some works try to explore deep learning based OA methods. VeeAlign (Iyer et al., 2021) is one of the representative methods, which utilizes word embeddings to predict the alignment.

Although some OA datasets and methods have been investigated and developed, at present there are no well-established benchmarks for event ontology alignment. In this paper, we propose EventOA, an event ontology alignment dataset, which can be used for understanding the events and evaluating the performance of systems analyzing the real world events.

Event Type Alignment				
Error Type	**Percent**	**Wikidata**	**FrameNet**	
			Reference	**Predict**
Ambiguity	54%	Treatment	Medical_intervention	Cure
Compound Word	28%	Deliberate_murder	Killing	Offenses
		Surgical_operation	Medical_intervention	Military_operation
Other	18%	-	-	-
Event Argument Alignment				
Spurious Correlation	42%	Mid-air_collision:Vehicle	Impact:Impactors	Impact:Speed
		Parade:Destination_point	Mass_motion:Goal	Mass_motion:Place
Same Category	34%	Stabbing:Perpetrator	Cause_harm:Agent	Cause_harm:Victim
		Performing_arts: Director	Performing_arts: Personnel	Performing_arts: Performer
Other	24%	-	-	-

Table 7: Error analysis. ":" is used to separate event type and event argument.

6 Conclusion

In this paper, we construct EventOA, an event ontology alignment dataset based on FrameNet and Wikidata. To overcome the challenges of the new task, we propose MEOA, which can utilize multi-view information to acquire richer and deeper semantic representations of events. Experiment results demonstrate that our MEOA method achieves better performance on EventOA and can serve as a strong baseline for future research.

Limitations

The limitations of our work are as follows: (1) As we construct EventOA for a new task by manually annotating, the data size can be further extended. (2) Our study is limited to English sources, and we hope work can pay attention to event ontologies in other languages. Building datasets for multilingual event ontology alignment would have a positive impact on applications in other languages beyond English.

Acknowledgements

We thank anonymous reviewers for their insightful comments and suggestions. This work is supported by the National Key Research and Development Program of China (No.2020AAA0106400), the National Natural Science Foundation of China (No.61976211, 6217625). This work is also supported by the Strategic Priority Research Program of Chinese Academy of Sciences (Grant No.XDA27020100), the Youth Innovation Promotion Association CAS, and Yunnan Provincial Major Science and Technology Special Plan Projects (No.202202AD080004).

References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The berkeley framenet project. In *Proceedings of COLING*, pages 86–90.

Olivier Bodenreider, Terry F Hayamizu, Martin Ringwald, Sherri De Coronado, and Songmao Zhang. 2005. Of mice and men: Aligning mouse and human anatomies. In *Proceedings of AMIA Annual Symposium*, pages 61–65.

Susan Brown, Claire Bonial, Leo Obrst, and Martha Palmer. 2017. The rich event ontology. In *Proceedings of EventStory workshop*, pages 87–97.

George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie M Strassel, and Ralph M Weischedel. 2004. The automatic content extraction (ace) program-tasks, data, and evaluation. In *Proceedings of LREC*, pages 837–840.

Zlatan Dragisic, Valentina Ivanova, Huanyu Li, and Patrick Lambrix. 2017. Experiences from the anatomy track in the ontology alignment evaluation initiative. *Journal of Biomedical Semantics*, 8:56.

Fredo Erxleben, Michael Günther, Markus Krötzsch, Julian Mendez, and Denny Vrandečić. 2014. Introducing wikidata to the linked data web. In *Proceedings of ISWC*, pages 50–65.

Daniel Faria, Catia Pesquita, Emanuel Santos, Matteo Palmonari, Isabel Cruz, and Francisco Couto. 2013. The agreementmakerlight ontology matching system. In *Proceedings of OTM*, pages 527–541.

Charles J. Fillmore. 1976. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences*, 280(1):20–32.

Simon Gottschalk and Elena Demidova. 2019. Eventkg–the hub of event knowledge on the web–and biographical timeline generation. *Semantic Web*, 10(6):1039–1070.

Yong Guan, Shaoru Guo, Ru Li, Xiaoli Li, and Hongye Tan. 2021. Frame semantic-enhanced sentence modeling for sentence-level extractive text summarization. In *Proceedings of EMNLP*, pages 4045–4052.

Shaoru Guo, Ru Li, Hongye Tan, Xiaoli Li, Yong Guan, Hongyan Zhao, and Yueping Zhang. 2020. A frame-based sentence representation for machine reading comprehension. In *Proceedings of ACL*, pages 891–896.

Yuan He, Jiaoyan Chen, Hang Dong, Ernesto Jiménez-Ruiz, Ali Hadian, and Ian Horrocks. 2022. Machine learning-friendly biomedical datasets for equivalence and subsumption ontology matching. In *Proceedings of ISWC*, pages 575–591.

Vivek Iyer, Arvind Agarwal, and Harshit Kumar. 2021. VeeAlign: Multifaceted context representation using dual attention for ontology alignment. In *Proceedings of EMNLP*, pages 10780–10792.

Heng Ji and Ralph Grishman. 2008. Refining event extraction through cross-document inference. In *Proceedings of ACL-HLT*, pages 254–262.

Ernesto Jiménez-Ruiz and Bernardo Cuenca Grau. 2011. Logmap: Logic-based and scalable ontology matching. In *Proceedings of ISWC*, pages 273–288.

Ernesto Jiménez-Ruiz, Bernardo Cuenca Grau, and Ian Horrocks. 2013. Logmap and logmaplt results for oaei 2013. In *Proceedings of OM@ISWC*, page 131–138.

Ernesto Jiménez-Ruiz, Asan Agibetov, Jiaoyan Chen, Matthias Samwald, and Valerie Cross. 2020. Dividing the ontology alignment task with semantic embeddings and logic-based modules. In *Proceedings of ECAI*, pages 784–791.

Naouel Karam, Abderrahmane Khiat, Alsayed Algergawy, Melanie Sattler, Claus Weiland, and Marco Schmidt. 2020. Matching biodiversity and ecology ontologies: challenges and evaluation results. *The Knowledge Engineering Review*, 35:e9.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *Proceedings of ICLR*.

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2007. A large-scale classification of english verbs. *Language Resources and Evaluation*, 42:21–40.

Prodromos Kolyvakis, Alexandros Kalousis, and Dimitris Kiritsis. 2018. DeepAlignment: Unsupervised ontology matching with refined word vectors. In *Proceedings of NAACL-HLT*, pages 787–798.

Li Li and Yun Yang. 2008. Agent negotiation based ontology refinement process and mechanisms for service applications. *Service Oriented Computing and Applications*, 2(1):15–25.

Wenjie Li, Mingli Wu, Qin Lu, Wei Xu, and Chunfa Yuan. 2006. Extractive summarization using inter- and intra- event relevance. In *Proceedings of ACL*, pages 369–376.

Vanessa Lopez, Victoria Uren, Marta Sabou, and Enrico Motta. 2011. Is question answering fit for the semantic web? a survey. *Semantic Web*, 2(2):125–155.

Thomas Pellissier Tanon, Gerhard Weikum, and Fabian Suchanek. 2020. Yago 4: A reason-able knowledge base. In *Proceedings of ESWC*, pages 583–596.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of EMNLP*, pages 1532–1543.

Jan Portisch, Michael Hladik, and Heiko Paulheim. 2019. Wiktionary matcher. In *Proceedings of OM@ISWC*, pages 181–188.

M. Pour, A. Algergawy, P. Buche, L. J. Castro, J. Chen, H. Dong, O. Fallatah, D. Faria, I. Fundulaki, S. Hertling, Y. He, I. Horrocks, M. Huschka, L. Ibanescu, E. Jimenez-Ruiz, N. Karam, A. Laadhar, P. Lambrix, H. Li, Y. Li, F. Michel, E. Nasr, H. Paulheim, C. Pesquita, T. Saveta, P. Shvaiko, C. Trojahn, C. Verhey, M. Wu, B. Yaman, O. Zamazal, and L. Zhou. 2022. Results of the ontology alignment evaluation initiative 2022. In *Proceedings of CEUR Workshop*, pages 84–128.

Jacobo Rouces, Gerard de Melo, and Katja Hose. 2015. Framebase: Representing n-ary relations using semantic frames. In *Proceedings of ESWC*, pages 505–521.

Vojtěch Svátek and Petr Berka. 2005. Ontofarm: Towards an experimental collection of parallel ontologies. In *Proceedings of ISWC*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of NIPS*, page 6000–6010.

Daya C. Wimalasuriya and Dejing Dou. 2010. Ontology-based information extraction: An introduction and a survey of current approaches. *Journal of Information Science*, 36(3):306–323.

Ondřej Zamazal and Vojtěch Svátek. 2017. The ten-year ontofarm and its fertilization within the onto-sphere. *Journal of Web Semantics*, 43:46–53.

A Examples of Event Ontology Alignment

We present the alignments of the event type and event argument in Table 8 and 9 to help understand event ontology alignment construction process.

In human annotation, the salary is determined by the average time of annotation and local labor compensation standard.

	Wikidata	FrameNet
Type Alignment	**War**	**Hostile_encounter**
Argument Alignment	Participant	Sides
	Conflict	Issue
	Duration	Duration
	Uses	Instrument
	Location Located_in/on_physical_feature Valid_in_place Located_in_protected_area	Place
	Time_period Inception Point_in_time Start_period End_period Start_time End_time	Time
	Has_effect Immediate_cause_of	Result
	Has_cause End_cause Has_contributing_factor Has_immediate_cause	Explanation
Argument without Alignment	- - - - - - - -	Side_1 Side_2 Degree Depictive Manner Means Particular_iteration Purpose
	Destroyed Commanded_by Number_of_injured Number_of_casualties Winner Victory Ordered_by	- - - - - - -

Table 8: Example of event type and event argument alignment of "*War*" and "*Hostile_encounter*". "-" represent arguments do not have correspondences.

	Wikidata	FrameNet
Type Alignment	**Assault**	**Attack**
Argument Alignment	Perpetrator Participant	Assailant
	Victim Target	Victim
	Point_in_time Start_time	Time
	Armament	Weapon
	Location	Place
	Has_cause	Explanation
Argument without Alignment	- - - - -	Duration Frequency Means Purpose Result
	Number_of_perpetrators Number_of_injured Number_of_deaths number_of_arrests	- - - -

Table 9: Example of event type and event argument alignment of “*Assault*” and “*Attack*”. “-” represent arguments do not have correspondences.

B SPARQL Query to Collect Wikidata-based Ontology

```
PREFIX bd: <http://www.bigdata.com/rdf#>
PREFIX wd: <http://www.wikidata.org/entity/>
PREFIX wikibase: <http://wikiba.se/ontology#>
PREFIX p: <http://www.wikidata.org/prop/>
PREFIX wdt: <http://www.wikidata.org/prop/direct/>

SELECT ?wikidata_event WHERE {
  ?wikidata_event wdt:P279* wd:Q1190554.
  SERVICE wikibase:label { bd:serviceParam
  wikibase:language "[AUTO_LANGUAGE],en". }}
```

Figure 4: The SPARQL query to obtain all subclasses of Wikidata’s “occurrence”.

C Event Argument Alignment Prediction

We compute the confidence score of event argument alignment S_a by taking similarity between fe in FrameNet and p in Wikidata.

$$S_a(fe_i, p_j) = cos\_sim(fe_i, p_j) \qquad (10)$$

We further use mean squared error as the loss to train our model (Iyer et al., 2021):

$$\mathcal{L}_a = \frac{1}{A}\sum(S_a(fe_i, p_j) - G_a(fe_i, p_j))^2 \quad (11)$$

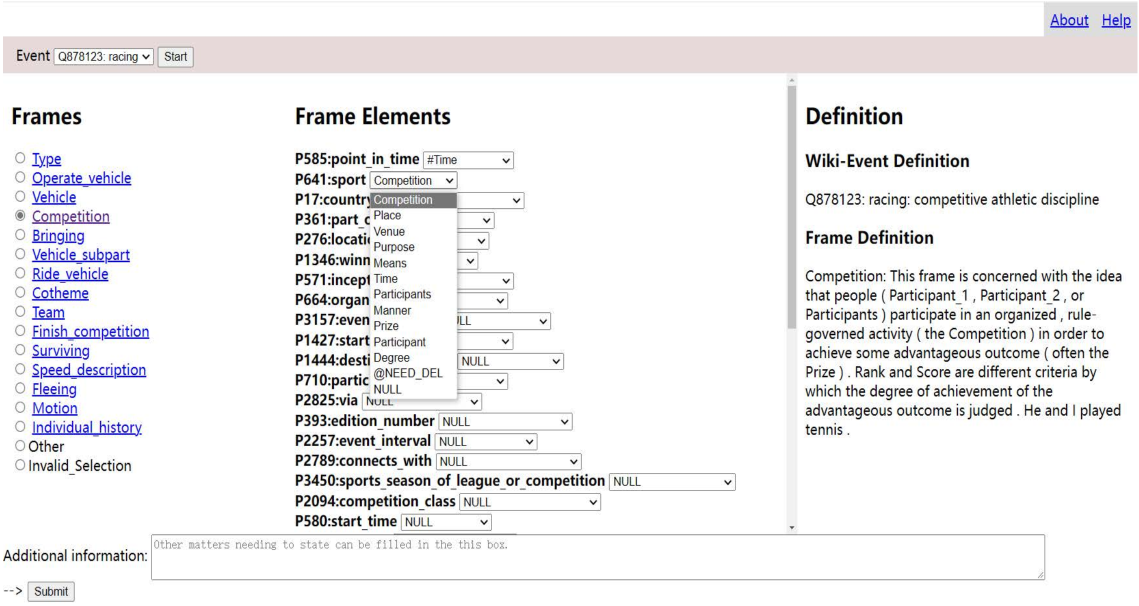
Where $S_a(\cdot)$ denotes the confidence score of event argument alignment, and $G_a(\cdot)$ denotes the ground truth label which is 1 if $fe_i \equiv p_j$ and 0 otherwise.

D Implementation Details and Baselines

We compare MEOA with various baselines. Specifically, AML (Faria et al., 2013) mixes various string-based matching methods to calculate matching scores. LogMap (Jiménez-Ruiz and Cuenca Grau, 2011; Jiménez-Ruiz et al., 2020) starts with a set of anchor mappings obtained from lexical comparison, then alternates between mapping repair and mapping discovery. Wiktionary (Portisch et al., 2019) is another top performing OA system that relies on the Wiktionary knowledge base. Word2Vec uses the vectors of their names and aliases to discover alignments. VeeAlign (Iyer et al., 2021) uses dual-attention mechanism to determine similarity between two ontologies.

We fine-tune MEOA for 6 epochs with a batch size of 48, and evaluated on the validation set for every 0.1 epoch, through which the best checkpoint is selected for prediction. The learning rate is set to 2e-5, while the loss functions used for EventOA is the mean squared error loss. The training uses a single NVIDIA GeForce RTX 3090 GPU.

E Annotation Interface for Human Annotators



Event Q878123: racing

Frames

○ Type
○ Operate_vehicle
○ Vehicle
● Competition
○ Bringing
○ Vehicle_subpart
○ Ride_vehicle
○ Cotheme
○ Team
○ Finish_competition
○ Surviving
○ Speed_description
○ Fleeing
○ Motion
○ Individual_history
○ Other
○ Invalid_Selection

Frame Elements

P585:point_in_time #Time
P641:sport Competition
P17:country
P361:part_c
P276:locati
P1346:winn
P571:incept
P664:organ
P3157:even ILL
P1427:start
P1444:desti NULL
P710:partic
P2825:via NULL
P393:edition_number NULL
P2257:event_interval NULL
P2789:connects_with NULL
P3450:sports_season_of_league_or_competition NULL
P2094:competition_class NULL
P580:start_time NULL

Definition

Wiki-Event Definition

Q878123: racing: competitive athletic discipline

Frame Definition

Competition: This frame is concerned with the idea that people (Participant_1 , Participant_2 , or Participants) participate in an organized , rule-governed activity (the Competition) in order to achieve some advantageous outcome (often the Prize) . Rank and Score are different criteria by which the degree of achievement of the advantageous outcome is judged . He and I played tennis .

Additional information: Other matters needing to state can be filled in the this box.

-->

Figure 5: A screenshot of the annotation interface for human annotators.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 6

☑ A2. Did you discuss any potential risks of your work?
Section 4.6

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Section Abstract, Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 2, Section 3

☑ B1. Did you cite the creators of artifacts you used?
Section 2, Section 3

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Section 2, Section 3

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 2, Section 3

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Section 2

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 2

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 2.4, Section 4.2

C ☑ Did you run computational experiments?

Section 4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section Appendix E

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section Appendix E

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 4.3, Section 4.4, Section 4.5

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 4.2

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
Section 2.4

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Section Appendix B

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Section 2.4

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Section 1, Section 2.1, Section 2.2

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Section 2

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Section 2.1, Section 2.2