GENEVA: Pushing the Limit of Generalizability for Event Argument Extraction with 100+ Event Types

Anonymous ACL submission

Abstract

Event Argument Extraction (EAE) deals with the task of extracting event-specific information from texts. EAE models usually require a large amount of annotated data for training, but procuring annotations is expensive for each new event type. To cater to the emerging event types and new domains in a realistic setting, it is growingly imperative for EAE models to be generalizable. However, most existing EAE benchmark datasets like ACE and ERE have limited diversity and coverage in terms of event types and cannot adequately evaluate the generalizability of EAE models. To alleviate this issue, we introduce GENEVA, a new dataset covering a diverse range of 115 event types and 187 argument role types. We create four benchmarking test suites in GENEVA to assess EAE models' generalizability. Additionally, we propose a new model AutoDEGREE which establishes a strong benchmark on these test suites. Lastly, we evaluate the generalizability of recent EAE systems from different model families and analyze their behaviors on GENEVA.[1]

1 Introduction

Event Argument Extraction (EAE) aims at extracting structured information of event-specific arguments and their roles for events from a pre-defined taxonomy. EAE has been studied for a long time (Sundheim, 1992; Grishman and Sundheim, 1996) and has been elemental in a wide range of applications like building knowledge graphs (Zhang et al., 2020), question answering (Berant et al., 2014), and various other NLP applications (Hogenboom et al., 2016; Yang et al., 2019b).

Previous works usually assume the availability of extensive and high-quality human annotations for training EAE models. However, in practice, there are a wide range of diverse events which usually have zero or few annotations as procuring annotations is an expensive process (Zhang et al., 2021). Hence, recent works focusing on generalizable EAE have gained more interests (Huang et al., 2018; Lyu et al., 2021; Sainz et al., 2022). These works utilize existing EAE datasets like ACE (Doddington et al., 2004) and ERE (Song et al., 2015) to verify the generalizability of the proposed models. However, as we show in Figure 1, these datasets have limited diversity and focus only on specific abstract event types. This limited diversity and reduced coverage restricts the ability of the existing datasets to more robustly evaluate the generalizability of EAE models.

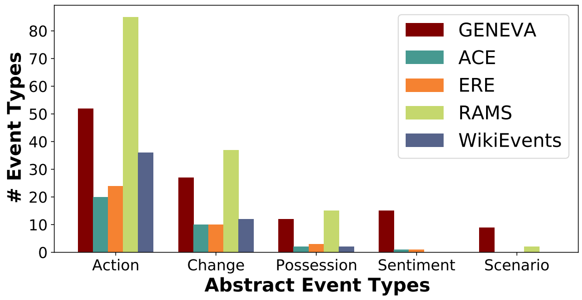

Figure 1: Distribution of event types into various abstractions for GENEVA, ACE, ERE, RAMS, and WikiEvents datasets. We observe that GENEVA is relatively more diverse in event type coverage. Abstract event types are defined as the top nodes of the event ontology tree created by MAVEN (Wang et al., 2020).

Towards this end, we introduce GENEVA (**G**eneralizability **BEN**chmarking Dataset for **EV**ent **A**rgument Extraction), a new diverse event argument extraction dataset covering a broad range of 115 event types spanning various abstract event types (Figure 1) and 187 argument roles to evaluate the generalizability of EAE models. GENEVA is repurposed from an existing semantic role labeling dataset, FrameNet (Baker et al., 1998), with manual selective filtering and merging. In order to test the models' ability to learn from limited training data and generalize to unseen event types, we design four benchmarking test suites. These test suites are distinctly different based on the training and test

[1] We will release our dataset and code upon acceptance.

data creation – (1) low resource, (2) few-shot, (3) zero-shot, and (4) cross-type transfer settings.

With the goal of pushing the limit of generalizability for EAE, we introduce a new model AutoDEGREE which inherits the current state-of-the-art EAE model in low-resource regime — DEGREE (Hsu et al., 2022). Like DEGREE, AutoDEGREE performs EAE via generating sentences that summarize all the event argument information using automated natural language prompts. On the other hand, AutoDEGREE enhances generalizability by introducing automated refinements to eliminate the human effort required for scaling up DEGREE to a wide range of events. We evaluate AutoDEGREE along with various other EAE models on the GENEVA test suites and demonstrate that AutoDEGREE establishes a strong generalizability benchmark on these test suites.

To sum up, we make the following contributions: (1) We introduce a new diverse EAE dataset GENEVA and design four benchmarking test suites to test the different aspects of generalizability of EAE models. (2) We introduce AutoDEGREE, a new EAE model which serves as a strong benchmark for the test suites in GENEVA. (3) We conduct a thorough evaluation of various EAE models on the test suites in GENEVA and show superior generalizability of generation-based models over classification-based models.

2 Related Work

Event Extraction Datasets: ACE (Doddington et al., 2004) is one of the earliest and most used Event Extraction datasets. The ACE event schema is further simplified and extended to ERE (Song et al., 2015). ERE was later used to create various TAC KBP Challenges (Ellis et al., 2014, 2015; Getman et al., 2017). These datasets cover only a limited amount of event types and argument roles, and thus, can't be utilized to adequately evaluate the generalizability of EAE models. MAVEN (Wang et al., 2020) introduced a massive and diverse dataset spanning a wide range of event types. However, the applicability of this dataset is limited to the task of Event Detection[2] and it does not contain argument role annotations. Recent works have introduced datasets like RAMS (Ebner et al., 2020), WikiEvents (Li et al., 2021), and DocEE (Tong et al., 2022); but the diversity of these datasets is restrictive to specific event categories as shown in Figure 1. Furthermore, these datasets are built with a focus on document-level event extraction task, while we target on evaluating generalizability of EAE models in sentence-level.

[2]Event Detection aims at only identifying the event type documented in the sentence.

Event Argument Extraction Models: Traditionally, EAE has been formulated as a classification problem (Nguyen et al., 2016). Previous classification-based approaches have utilized pipelined approaches (Yang et al., 2019a; Wadden et al., 2019) as well as incorporating global features for joint inference (Li et al., 2013; Yang and Mitchell, 2016; Lin et al., 2020). However, most of these classification approaches are data-hungry and do not generalize well in the low-data setting (Liu et al., 2020; Hsu et al., 2022). To improve generalizability, some works have explored better usage of label semantics by formulating EAE as a question-answering task (Liu et al., 2020; Li et al., 2020; Du and Cardie, 2020). Recent approaches have explored the use of natural language generative models for classification and structured prediction for better generalizability (Schick and Schütze, 2021a,b). TANL (Paolini et al., 2021) treats EAE as a translation between augmented languages, whereas Bart-Gen (Li et al., 2021) is another generative approach that focuses on document-level EAE. DEGREE (Hsu et al., 2022) is a recently introduced state-of-the-art generative model which has shown better performance in the limited data regime. Another set of works transfer knowledge from similar tasks like abstract meaning representation and semantic role labeling (Huang et al., 2018; Lyu et al., 2021; Zhang et al., 2021) to perform EAE.

Since the evaluation of these models is done on previous EAE datasets, it is unclear if these approaches can be generalized to handle a diverse set of events. In our work, we benchmark various classes of previous models on our benchmarking test suites. Furthermore, we propose a new model AutoDEGREE which outperforms previous models and serves as a strong baseline for future works.

3 GENEVA Dataset

Annotating data for EAE for a diverse set of events is a resource-heavy and expensive process. Rather, we take advantage of the shared properties between Semantic Role Labeling (SRL) and EAE and utilize an existing dataset FrameNet to create a wide-coverage dataset for EAE. We follow the event

Frame	Arrest	Visiting	Travel
Frame Elements	Authorities, Charges, Offense, Suspect, *Co-participant, Time, Means, Place, Purpose, Type, Source of legal authority, Manner*	Agent, Entity, *Frequency, Depictive, Duration, Means, Iterations, Manner, Purpose, Normal location, Place, Dependent state, Time*	Traveler, Path, Source, Goal, Direction, Mode of Transportation, Area, *Explanation, Frequency, Baggage, Depictive, Iterations, Co-participant, Duration, Manner, Speed, Time, Descriptor, Period of iterations, Distance, Means, Purpose, Result*

Table 1: An illustration of the complex structure for different frames from FrameNet. During GENEVAcreation, frame elements in the same color are merged into a single argument role, while those in *italics* are filtered out.

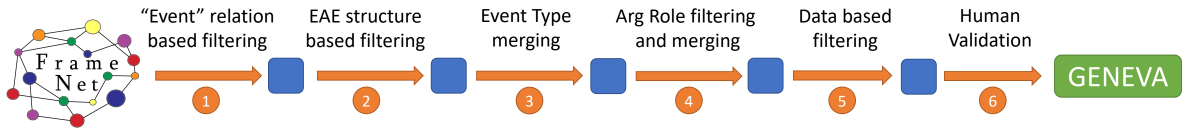

Figure 2: Figure highlighting the various operations performed to create our proposed EAE dataset GENEVA from the SRL dataset FrameNet.

definition from ACE (Doddington et al., 2004) (described in Appendix A). Here, we focus on discussing our data creation process.

3.1 FrameNet for EAE

SRL and EAE are similar tasks in that SRL assigns semantic roles to phrases in the sentence and EAE aims at extracting event-specific argument roles from the sentence. Owing to these similarities, we utilize FrameNet (Baker et al., 1998)[3] - a comprehensive SRL dataset comprising of 1200+ semantic frames (Fillmore et al., 1976) - to create an EAE dataset. The definition of a *frame* is rather loose and can be understood as the holistic background that unites similar words.[4] Each frame is annotated with frame-specific semantic roles (*frame elements*) and words that evoke the frame (*lexical units*). We utilize FrameNet for EAE by mapping selective *frames* that have "Event" relations as events. Correspondingly, we can map the *lexical units* as event triggers and *frame elements* as argument roles. For example in Table 1, the frame *Arrest* and its frame elements can be mapped to the event *Arrest-Jail* of the ACE dataset and its argument roles respectively.

However, the applicability of FrameNet for EAE has been limited. This is primarily because FrameNet prioritizes lexicographic and linguistic completeness (Aguilar et al., 2014), while EAE is a higher-level task requiring extraction of distinct and succinct information. This difference leads to two major challenges in using FrameNet for EAE: (1) *FrameNet frames are too fine-grained and many times indistinguishable from the aspect of EAE*, and (2) *FrameNet frames have a complex structure comprising of a wide range of frame elements which may all not be relevant for EAE.*

We provide an example of these challenges in Table 1 where we show two distinct frames from FrameNet - *Visiting* and *Travel*. However, from the perspective of EAE, these frames are similar and can be merged into a single event (first challenge). Furthermore, we observe that these frames have a wide range of frame elements many of which are rarely used (e.g. *Periods of iteration*) while some of them are quite generic (eg. *Manner*). Only a partial portion of these frame elements are indeed relevant for EAE (highlighted in non-italics in Table 1) which demonstrates the second challenge.

3.2 Creation of GENEVA

In order to bridge the differences between the task definitions of SRL and EAE (discussed in Section 3.1), we perform several merging and filtering operations to create more distinctive and representative event types and argument roles. We also perform a human validation to ensure the significance of these operations. We show the transformation of FrameNet into GENEVA in Figure 2 and describe each of these operations in detail below.

Event Filtering: This operation deals with filtering of frames from FrameNet which are relevant for the task of EAE. The first set of filtering is done by selecting frames which have a relation with the "Event" frame inspired by Li et al. (2019) and leads to a total of 289 frames (Step 1 in Figure 2). Next, we use the structure of events and filter out frames which do not have any arguments or datapoints

[3] FrameNet Data Release 1.7 by http://framenet.icsi.berkeley.edu is licensed under a Creative Commons Attribution 3.0 Unported License.

[4] www.web.stanford.edu/~jurafsky/slp3/19.pdf

Dataset	#Sentences	#Event Types	#Arg Types	#Event Mentions	#Arg Mentions	Avg. Event Mentions	Avg. Arg Mentions
ACE	18,927	33	22	5,055	6,040	153.18	274.55
ERE	17,108	38	21	7,284	10,479	191.68	499
GENEVA	3,673	115	187	7,576	11,163	65.88	59.7

Table 2: Statistics for the different datasets for Event Argument Extraction. The third and fourth columns indicate the unique number of event types and argument roles. The fifth and sixth column are the number of event and argument mentions in the dataset. The last two columns indicate the average number of mentions per event and argument role.

(Step 2). This yields a total of 230 frames.

Event Merging: This operation deals with merging similar frames into a single event type (e.g. *Visiting* and *Travel*). Following their hierarchical event ontology from MAVEN (Wang et al., 2020) we manually merge similar and fine-grained frames to reduce the total number of event types to 158 (Step 3), covering upto 36% annotated sentences of the FrameNet dataset.

Argument Role Filtering and Merging: Each frame comprises of a large set of frame elements in FrameNet. At this step, we aim to filter and merge them into a reduced set of argument roles (Step 4). We filter frame elements with high precision by removing all the *non-core* frame elements as they are generic and not frame-specific by definition (highlighted in gray in Table 1). We further remove all argument roles with no mentions in the data. To facilitate better overlap of argument roles across events and reduce redundancy, we manually merge frame elements (e.g. *Agent* in *Visiting* frame and *Traveler* in *Travel* frame) based on their relevance and similarity to each other. This yields us with a total of 250 argument roles.

Data Based Filtering: We set a minimum data requirement to 5 event mentions (in order to aid better evaluation) and remove event types that do not meet that criteria (e.g. *Lighting*). The final event schema of GENEVA comprises of 115 event types and 187 argument roles. We also organize our events into the hierarchical event schema devised by MAVEN (shown in Appendix D).

Human Validation: In order to distinguish GENEVA from FrameNet and validate the utility of our merging operations, we set up a human validation experiment (Step 6). We present the human annotators with three sentences - one primary and two candidates - and ask them if the event type described in the primary sentence is similar to the event types in either of the candidates or distinct from both (Example in Appendix H). One candidate is chosen as a sentence from one of the frames merged with the primary event, while the other candidate is chosen from a similar unmerged frame, which is a sibling event of the primary event discovered from the event ontology. The annotators chose the merged frame candidates on an average of 87%. The validation was done by three annotators over 61 sampled triplets and with 0.7 inter-annotator agreement measured in Fleiss' kappa (Fleiss, 1971). This human validation ensures high dataset quality as well as underlines the significance of the various operations performed for the creation of GENEVA.

3.3 Data Analysis

Here, we show how GENEVA is different from previous EAE datasets of ACE and ERE, and is more suited to evaluate the generalizability of EAE models. The major statistics for GENEVA are shown in Table 2 along with its comparison with ACE and ERE. We observe that GENEVA has fewer sentences compared to the other datasets. Nevertheless, it has thrice the number of event types and 8 times the number of argument roles relative to ACE/ERE. Furthermore, the number of event and argument role mentions are more compared to the previous datasets. Naturally, the average number of mentions per event and argument role (refer to the last two columns in Table 2) is much lesser for GENEVA. We categorize the event types for GENEVA and ACE into abstract event types (as defined in MAVEN (Wang et al., 2020)) and show their distribution in Figure 1. The figure depicts how ACE events are concentrated only in specific abstractions of Action and Change, while GENEVA has a more diverse distribution. Overall, these statistics show how GENEVA is more diverse and challenging than the previous datasets.

Due to the high number of event types and argument roles, GENEVA is a highly dense dataset. We plot the distribution of argument roles per sentence[5] for ACE, ERE, and GENEVA in Figure 3.

[5] We remove no event mention sentences for ACE/ERE.

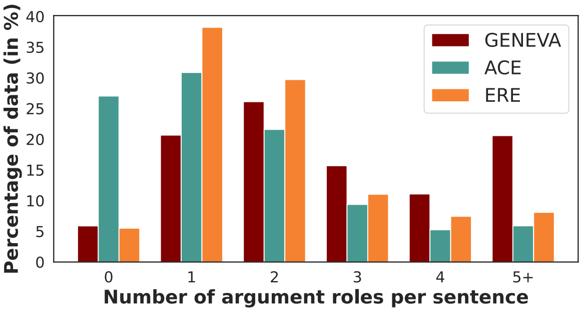

Figure 3: Argument roles per sentence as percentage of data for ACE, ERE and GENEVA datasets.

Both ACE and ERE have a high proportion of sentences ($> 70\%$) with up to 2 argument roles. In contrast, GENEVA is denser with almost 50% of sentences having 3 or more arguments and more than 20% of sentences with 5+ arguments.

3.4 Benchmarking Test Suites

With a focus on the evaluation of the generalizability of the EAE models, we fabricate four benchmarking test suites clubbed into two higher-level settings, as described below:

Limited Training Data: This setting mimics the realistic scenario when there are fewer annotations available for the target events and evaluates models' ability to learn from limited training data. We present two test suites for this setting:

- Low resource (LR): Training data is created by *randomly* sampling n event mentions.[6] We record the model performance across a spectrum from extremely low resource ($n = 1$) to moderately resource ($n = 1200$) settings.
- Few-shot (FS): Training data is curated by sampling n event mentions *uniformly* across all events. This sampling strategy avoids biases towards high data events and assesses the model's ability to perform well uniformly across events. We study the model performance from one-shot ($n = 1$) to five-shot ($n = 5$) for this test suite.

Unseen Event Data: The second setting focuses on the scenario when there is no annotation available for the target events. This helps test models' ability to generalize to unseen events and argument roles. We propose two test suites:

- Zero-shot (ZS): The top 10 events in terms of data availability is used to create the training data and the remaining 105 events are utilized for testing. Intending to study the impact of event diversity on the zero-shot model performance, we create three training datasets by sampling a fixed 450 sentences[7] for m events from the larger training corpus. We vary m from 1 most-frequent event to 10 events.
- Cross-type Transfer (CTT): Adhering to the hierarchical event schema (refer to Appendix D), we curate a training dataset comprising of events of a single abstraction category (e.g. Scenario), while the test dataset comprises of events of all other abstraction types. This test suite also assesses models' transfer learning strength.

We report the data statistics for these benchmarking setups in Appendix B. For each of the test suites involving sampling, we sample 5 different datasets[8] and report the average model performance to account for the sampling variation.

4 Proposed Model — AutoDEGREE

In our work, we introduce a new model AutoDEGREE which aims to provide better generalizability for EAE. AutoDEGREE reforms a recent approach DEGREE (Hsu et al., 2022) with automated refinements. These refinements aid AutoDEGREE to scale up robustly to a wide range of event types while eliminating the human effort requirements of DEGREE. In this section, we first briefly introduce the base DEGREE model and then describe our proposed model AutoDEGREE.

4.1 DEGREE

DEGREE[9] is an encoder-decoder based generative model which utilizes natural language templates as part of input prompts. The input prompt comprises of three components - (1) *Event Type Description* which provides a definition of the given event type, (2) *Query Trigger* which indicates the trigger word for the event mention, and (3) *EAE Template* which is a natural sentence combining the different argument roles of the event. Conditioned on the input prompt, the model generates a natural language sentence with the extracted arguments. Restructuring argument roles into natural language input prompts helps DEGREE better leverage label semantics, and

[6]Due to a high variation in the number of event mentions per sentence, a fixed number of sampled sentences could have a varied number of event mentions. To discount this variability, we create the sampled training data such that each of them has a fixed number of n event mentions.

[7]Fixing the training data size removes the confounding variable of data size for the study.

[8]All datasets will be released for reproducibility purpose.

[9]For our work, we consider the EAE version of DEGREE.

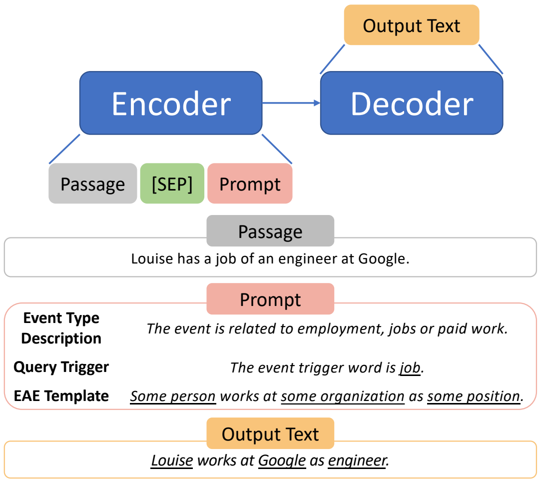

Figure 4: Model architecture of DEGREE (top half) and an illustration of a manually created prompt for the event type *Employment* (bottom half).

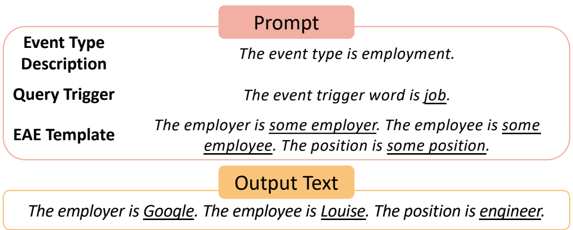

Figure 5: An illustration of an automatically generated prompt by AutoDEGREE for the event type *Employment*.

this fundamentally assists it to generalize in the low-data setting. We illustrate DEGREE along with an example of its input prompt design in Figure 4.

Despite the superior performance of DEGREE in the low-data setting, it can not be deployed on GENEVA. This is because DEGREE requires manual human effort for the creation of input prompts for each event type and argument role and can't be scaled to 115 event types and 187 argument roles in GENEVA. Thus, there is a need to automate the manual human effort to scale up DEGREE.

4.2 AutoDEGREE

AutoDEGREE exploits the same working principle of using natural language input prompts as DEGREE, while scaling up the prompt creation pipeline via automated refinements. DEGREE requires human effort for two input prompt components - (1) Event Type Description and (2) EAE Template. We describe the automated refinements in AutoDEGREE for these components below.

4.2.1 Automating Event Type Description

Event type description is a natural language sentence describing the event type. In order to automate this component, we propose a simple heuristic that creates a simple natural language sentence mentioning the event type - "*The event type is {event-type}*". , as illustrated in Figure 5.

4.2.2 Automating EAE Template

EAE template generation in DEGREE can be split into two subtasks, which we discuss in detail below.

Argument Role Mapping: This subtask maps each argument role to a natural language placeholder phrase based on the characteristics of the argument role. For example, the argument role *Employer* is mapped to "*some organization*" in Figure 4. While training, DEGREE learns to replace these placeholders in the prompt with the arguments from the passage. Mapping each unique argument role to a placeholder phrase requires commonsense knowledge, and thus rendering this subtask manual in nature.

For automating this mapping process, we propose a simple refinement of self mapping. Self mapping maps each argument role to a self-referencing placeholder phrase "*some {arg-name}*", where *{arg-name}* is the argument role itself. For example, the argument role *Employer* would be mapped to "*some employer*". We illustrate an example of this heuristic in Figure 5.

Template Generation: The second subtask requires generating a natural sentence(s) using the argument role mapped placeholder phrases (as shown in Figure 4). Each event type comprises of a distinct set of argument roles. Thus, generating EAE templates for each event type is tedious and created by human in DEGREE.

In order to automate this subtask, AutoDEGREE utilizes an event-agnostic template composed of argument role-specific sentences. For each argument role in the event, we generate a sentence of the form "*The {arg-name} is {arg-map}*." where *{arg-name}* and *{arg-map}* is the argument role and its mapped placeholder phrase respectively. For example, the sentence for argument role *Employer* with self mapping would be "*The employer is some employer.*". The final event-agnostic template is a simple concatenation of all the argument role sentences. We provide an illustration of the event-agnostic template in Figure 5.

5 Experimental Setup

In this section, we describe the baseline models and the evaluation metrics for our experiments.

5.1 Baseline Models

We aim to evaluate the generalizability of various representative EAE models on our GENEVA benchmarking test suites. These models include (1) **DyGIE++** (Wadden et al., 2019), a traditional classification based model utilizing multi-sentence BERT encodings and span graph propagation. (2) **OneIE** (Lin et al., 2020), a multi-tasking objective based model exploiting global features for optimization. (3) **BERT_QA** (Du and Cardie, 2020), a BERT-based model leveraging label semantics by framing EAE as a machine reading comprehension task. In order to scale BERT_QA to a wide range of argument roles, we generate question queries of the form "*What is {arg-name}?*" for each argument role *{arg-name}*. (4) **TANL** (Paolini et al., 2021), a language generation model which treats EAE as a translation task. We benchmark our proposed model AutoDEGREE with these baseline models.

5.2 Evaluation Metrics

Following the traditional evaluation for EAE tasks, we report the micro F1 scores for argument classification. To encourage better generalization across wide range of events, we also use macro F1 score that reports the average of F1 scores for each event type. For the limited data test suites, we record a model performance curve, wherein we plot the F1 scores against the number of training instances.

6 Results and Analysis

Following the benchmarking setups discussed in Section 3.4, we organize the main experimental results into limited training data and unseen event data settings. When trained on complete training data, we observe that OneIE achieves a poor micro F1 score of just 38.84 while all other models achieve F1 scores above 55. This can be attributed to the model design of OneIE as it is unable to handle overlapping argument roles.[10] Due to its inferior performance, we do not include OneIE in the benchmarking results.

6.1 Limited Training Data

Limited training data setting comprises of the low resource and the few-shot test suites. We present the model benchmarking results in terms of macro F1 and micro F1 scores for the low resource test suite in Figure 6. We observe that AutoDEGREE

[10]One key attribute of GENEVA is that arguments overlap with each other quite frequently in a sentence.

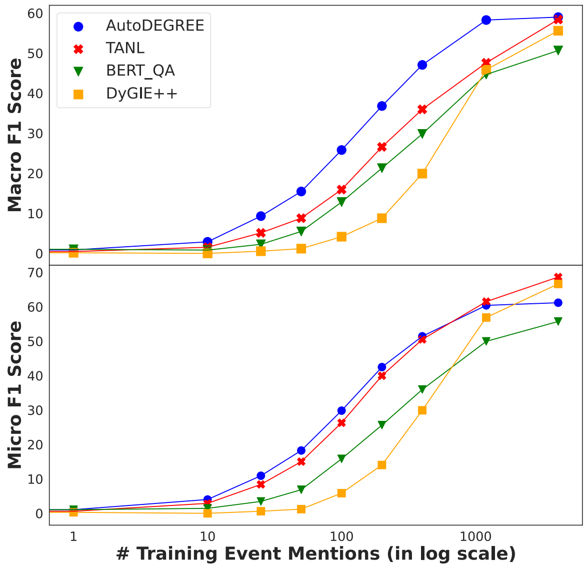

Figure 6: Model performance in macro F1 (top) and micro F1 (bottom) scores against the number of training event mentions (log-scale) for the low resource suite.

beats all other baselines significantly in terms of macro F1 and performs well uniformly across all event types. Although TANL and DyGIE++ achieve high micro F1 when trained on high number of training instances, their macro scores are still relatively poor. This indicates that these models are biased towards specific events and do not generalize well.

In Figure 7, we show the benchmarking results on the few-shot test suite. The results showcase the clear hierarchy of the model performance, wherein AutoDEGREE significantly outperforms all other models. We also observe the poor performance of traditional classification-based approaches like DyGIE++ and this underlines the importance of using label semantics for better generalizability.

6.2 Unseen Event Data

This data setting includes the zero-shot and the cross-type transfer test suites. We collate the results in terms of micro F1 scores for both the test suites in Table 3. Models like DyGIE++ and TANL cannot support unseen events or argument roles and thus, we do not include these models in the experiments for these test suites.

From Table 3, we observe that AutoDEGREE achieves the best score across all setups of the zero-shot setting. Furthermore, for the cross-type transfer, we observe that the AutoDEGREE outperforms BERT_QA by a significant margin of almost 20 F1 points. This establishes the superior generalizabil-

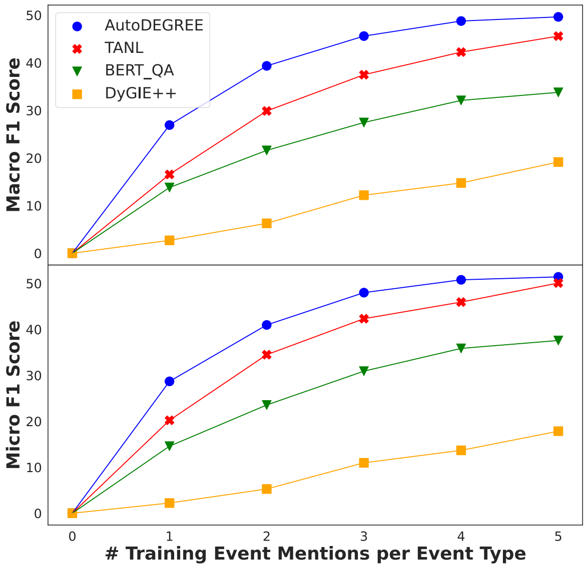

Figure 7: Model performance in macro F1 (top) and micro F1 (bottom) scores against the number of training event mentions per event for the few-shot suite.

Model	ZS-1	ZS-5	ZS-10	CTT
BERT_QA	5.21	23.15	23.23	7.83
AutoDEGREE	**13.91**	**33.06**	**35.47**	**27.26**

Table 3: Model performance in micro F1 scores for the zero-shot (ZS) and cross-type transfer (CTT) test suites. ZS-1, ZS-5, and ZS-10 indicate 1, 5, and 10 event types for training respectively. We exclude TANL and DyGIE++ as they cannot transfer to unseen events.

ity and transferability of AutoDEGREE to unseen event types and argument roles. We also record performance gains for both models as we increase the number of events in the training data. On the other hand, these gains reduce as the number of training events increases. Thus, we conclude that event diversity helps improve zero-shot performance but provides marginally reducing gains.

6.3 Case Study: Is ACE diverse enough?

In this section, we conduct a case study to analyze how the limited diversity of ACE can affect the generalizability of EAE models. We compare the performance of two models with different initializations - (1) AutoDEGREE pre-trained on the ACE dataset and (2) AutoDEGREE with no pre-training - on the zero-shot with 10 event types benchmarking setup. We dissect the F1 scores into different abstract event types and show the results in Table 4.

We observe that pre-training yields major improvements for the abstractions of Action, Possession, and Change - which are well-represented in ACE. On the other hand, we observe lower or negative performance improvement for the abstractions of Sentiment and Scenario - which are not represented in ACE. This trend clearly shows that the lack of diversity in ACE restricts the models' ability to generalize to out-of-domain event types. We also highlight the significance of GENEVA as its diverse evaluation setup helps analyze these trends.

Abstract Event Type	Scratch Model	Pre-Trained Model	Δ
Action	28.11	32.32	4.21
Possession	40.19	44.41	4.21
Change	41.15	44.92	3.77
Sentiment	43.39	44.92	1.53
Scenario	40.77	32.24	-8.53

Table 4: Model Performance in micro F1 on zero-shot with 10 event types split by abstract event types for (1) AutoDEGREE with no pre-training (Scratch Model), and (2) Pre-Trained AutoDEGREE on ACE (Pre-Trained Model). Δ: model performance difference.

6.4 Discussion

Overall, our experiments on the various benchmarking test suites reveal many insights. First, we observe the superior generalizability of AutoDEGREE. Second, macro score evaluation reveals how models like TANL and DyGIE++ are biased towards specific events and show poor generalization. Overall, we observe better performance of generation-based models, like TANL and AutoDEGREE compared to classification-based models, like OneIE and DyGIE++ across all test suites.

7 Conclusion and Future Work

In this paper, we introduce a new diverse EAE dataset GENEVA comprising of a wide range of event types and argument roles. We develop four benchmarking test suites for evaluating model generalizability on the dataset and benchmark various representative EAE models. We also propose AutoDEGREE which shows superior generalization across the different test suites. Future work includes expansion of this dataset to cover more diverse event types and argument roles. Efforts can also be taken to improve the automated heuristics for AutoDEGREE and in turn, pushing the limit of generalizability furthermore.

Limitations

We would like to highlight a few limitations of our work. First, we would like to point out that GENEVA is designed to evaluate the generalizability of EAE models. Although the dataset contains event type and event trigger annotations, it can only be viewed as a partially-annotated dataset if end-to-end event extraction is considered. Furthermore, there is no guarantee that all possible events in the sentence are exhaustively annotated. Finally, GENEVA is derived from an existing dataset FrameNet. Despite the exhaustive human efforts put into the argument selection and frame merging, the label quality in GENEVA is still influenced by the annotation quality of FrameNet.

Ethical Consideration

We would like to list a few ethical considerations for our work. First, GENEVA is derived from FrameNet which comprises of annotated sentences from various news articles. Many of these news articles cover various political issues which might be biased and sensitive to specific demographic groups. We encourage careful consideration for utilizing this data for applying trained models in this dataset for real-world production. Another consideration for our work would be concerning the applications of our proposed model AutoDEGREE, as it is a generative approach. Despite best efforts to exercise control over the output generation, it is not guaranteed to produce sentences that adhere to the template and are safe in nature. It can be susceptible to adversarial attacks and produce incoherent and unsafe sentences.

References

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.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, Montreal, Quebec, Canada. Association for Computational Linguistics.

Jonathan Berant, Vivek Srikumar, Pei-Chun Chen, Abby Vander Linden, Brittany Harding, Brad Huang, Peter Clark, and Christopher D. Manning. 2014. Modeling biological processes for reading comprehension. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1499–1510, Doha, Qatar. 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.

Seth Ebner, Patrick Xia, Ryan Culkin, Kyle Rawlins, and Benjamin Van Durme. 2020. Multi-sentence argument linking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8057–8077, Online. Association for Computational Linguistics.

Joe Ellis, Jeremy Getman, Dana Fore, Neil Kuster, Zhiyi Song, Ann Bies, and Stephanie M Strassel. 2015. Overview of linguistic resources for the tac kbp 2015 evaluations: Methodologies and results. In *TAC*.

Joe Ellis, Jeremy Getman, and Stephanie M Strassel. 2014. Overview of linguistic resources for the tac kbp 2014 evaluations: Planning, execution, and results. In *Proceedings of TAC KBP 2014 Workshop, National Institute of Standards and Technology*, pages 17–18.

Charles J Fillmore et al. 1976. Frame semantics and the nature of language. In *Annals of the New York Academy of Sciences: Conference on the origin and development of language and speech*, volume 280, pages 20–32. New York.

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

Jeremy Getman, Joe Ellis, Zhiyi Song, Jennifer Tracey, and Stephanie M Strassel. 2017. Overview of linguistic resources for the tac kbp 2017 evaluations: Methodologies and results. In *TAC*.

Ralph Grishman and Beth Sundheim. 1996. Message Understanding Conference- 6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Frederik Hogenboom, Flavius Frasincar, Uzay Kaymak, Franciska de Jong, and Emiel Caron. 2016. A survey of event extraction methods from text for decision support systems. *Decis. Support Syst.*, 85:12–22.

I-Hung Hsu, Kuan-Hao Huang, Elizabeth Boschee, Scott Miller, Prem Natarajan, Kai-Wei Chang, and Nanyun Peng. 2022. Degree: A data-efficient generative event extraction model. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

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.

Fayuan Li, Weihua Peng, Yuguang Chen, Quan Wang, Lu Pan, Yajuan Lyu, and Yong Zhu. 2020. 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.

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.

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.

Wei Li, Dezhi Cheng, Lei He, Yuanzhuo Wang, and Xiaolong Jin. 2019. Joint event extraction based on hierarchical event schemas from framenet. *IEEE Access*, 7:25001–25015.

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.

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.

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.

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cícero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. 2021. Structured prediction as translation between augmented natural languages. In *9th International Conference on Learning Representations (ICLR)*.

Oscar Sainz, Itziar Gonzalez-Dios, Oier Lopez de Lacalle, Bonan Min, and Eneko Agirre. 2022. Textual entailment for event argument extraction: Zero- and few-shot with multi-source learning. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2439–2455, Seattle, United States. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021a. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021b. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.

Zhiyi Song, Ann Bies, Stephanie M. 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, (EVENTS@HLP-NAACL)*.

Beth M. Sundheim. 1992. Overview of the fourth Message Understanding Evaluation and Conference. In *Fourth Message Uunderstanding Conference (MUC-4): Proceedings of a Conference Held in McLean, Virginia, June 16-18, 1992*.

MeiHan Tong, Bin Xu, Shuai Wang, Meihuan Han, Yixin Cao, Jiangqi Zhu, Siyu Chen, Lei Hou, and Juanzi Li. 2022. DocEE: A large-scale and fine-grained benchmark for document-level event extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3970–3982, Seattle, United States. 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, Wangyi Jiang, Rong Han, Zhiyuan Liu, Juanzi Li, Peng Li, Yankai Lin, and Jie Zhou. 2020. MAVEN: A Massive General Domain Event Detection Dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1652–1671, Online. Association for Computational Linguistics.

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. 2019a. 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.

Yang Yang, Deyu Zhou, Yulan He, and Meng Zhang. 2019b. Interpretable relevant emotion ranking with event-driven attention. 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 177–187, Hong Kong, China. Association for Computational Linguistics.

Hongming Zhang, Xin Liu, Haojie Pan, Yangqiu Song, and Cane Wing-Ki Leung. 2020. *ASER: A Large-Scale Eventuality Knowledge Graph*, page 201–211. Association for Computing Machinery, New York, NY, USA.

Hongming Zhang, Haoyu Wang, and Dan Roth. 2021. Zero-shot Label-aware Event Trigger and Argument Classification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1331–1340, Online. Association for Computational Linguistics.

A Task Definition

An *event* is a specific occurrence involving multiple participants and is labeled with a specific *event type*. An *event mention* is a sentence in which the event is described. An *event trigger* is a word phrase which best expresses the event occurrence in an event mention. An *event argument* is a word phrase that mentions an event-specific attribute or participant and is labeled with a specific *argument role*. EAE aims at identifying event arguments in event mentions and classifying them into argument roles. EAE models can utilize event type and the associated event trigger as additional information for the task. For example, in the illustration in Figure 8, EAE requires the extraction of the argument roles of *Helper*, *Benefiter*, and *Goal* using the event type *Assistance* and the event trigger *helping* (highlighted in blue).

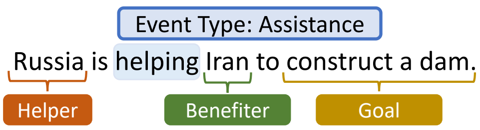

Figure 8: An illustration of Event Argument Extraction for the Assistance event type, which comprises of argument roles like Helper, Benefiter, and Goal.

B Data Statistics for different benchmarking test suites

We show the data statistics for the various benchmarking scenarios in Table 5. For the training set of the low resource and few-shot scenarios (indicated by $*$ in Table 5), we sample a smaller training set (as discussed in Section 3.4). For the zero-shot setup, the top 10 event types contribute to a large pool of $1,889$ sentences. For the test suites, a fixed number of 450 and 115 sentences are sampled for the training and the development set (indicated by $+$ in Table 5) from this larger pool of data.

C Event Type Distribution for GENEVA

We show the distribution of event mentions per event type for GENEVA in Figure 9. We observe a highly skewed distribution with 44 event types having less than 25 event mentions. Furthermore, 93 event types have less than 100 event mentions. We believe that this resembles a more practical scenario where there is a wide range of events with limited event mentions while a few events have a large number of mentions.

	LR/FS	ZS	CTT
# Train Sentences	$1,967^*$	450^+	268
# Dev Sentences	778	115^+	66
# Test Sentences	928	$1,784$	$3,339$

Table 5: Data statistics of the number of test sentences for the different benchmarking test suites. Here, LR: Low Resource, FS: Few-shot, ZS: Zero-shot, CTT: Cross-Type Transfer. $*$ and $+$ indicate that certain sampling is done for creating these datasets. More details are provided in the text.

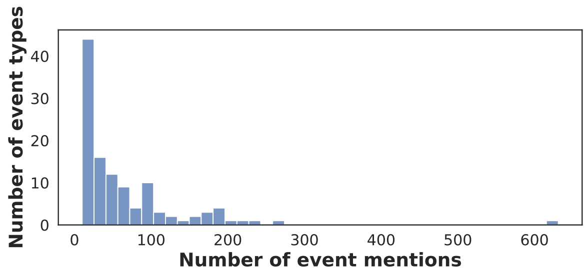

Figure 9: Distribution of event types by the number of event mentions in GENEVA.

D Event Schema Organization for GENEVA

The broad set of event types in GENEVA can be organized into a hierarchical structure based on event type abstractions. Adhering to the hierarchical tree structure introduced in MAVEN, we show the corresponding organization for event types in GENEVA in Figure 12. The organization mainly assumes five abstract event categories - Action, Change, Scenario, Sentiment, and Possession. The most populous abstract type is Action with a total of 53 events, while Scenario abstraction has the lowest number of 9 events.

We also study the distribution of event mentions per event type in Figure 12 where the bar heights are indicative of the number of event mentions for the corresponding event type (heights in log-scale). We observe that the most populous event is *Statement* which falls under the Action abstraction. On the other hand, the least populous event is *Recovering* which belongs to the Change abstraction.

GENEVA comprises of a diverse set of 115 event types and it naturally shares some of these with the ACE dataset. In Figure 12, we show the extent of the overlap of the mapped ACE events in the GENEVA event schema (text labels colored in red).[11] We can observe that although there is some

[11] We only show the events that could be directly mapped

	DEGREE	AutoDEGREE
ACE Dataset	73.5	72.7

Table 6: Model Performance in terms of F1 score for DEGREE and AutoDEGREE on the ACE dataset.

Model	ZS-1	ZS-5	ZS-10	CTT
BERT_QA	3.12	23.15	19.99	19.93
AutoDEGREE	**12.61**	**32.29**	**34.8**	**27.27**

Table 7: Model performance in macro F1 scores for the zero-shot (ZS) and cross-type transfer (CTT) test suites. ZS-1, ZS-5, and ZS-10 indicate the test suites with 1, 5, and 10 event types for training. We exclude TANL and DyGIE++ from the results as they cannot transfer to unseen events.

overlap between the datasets, GENEVA brings in a vast pool of new event types. Furthermore, most of the overlap is for the Possession and Action abstraction types, while very few or none of the overlaps fall in the Sentiment and Scenario abstraction types.

E Comparison of AutoDEGREE with DEGREE

In our work, we introduce a new model AutoDEGREE which provides automated and scaling refinements over the DEGREE model. Here, we provide a comparison of these two models and a correpsonding ablation study for the various componenets of the AutoDEGREE model. We train the AutoDEGREE on the standard ACE dataset and show the results in Table 6.

F Macro F1 Results for Unseen Event Data

The unseen event data setting comprises of the zero-shot and the cross-type transfer test suites. We present the results for model performance for these test suites in terms of macro F1 scores in Table 7. We observe similar trends like observed for micro F1 scores wherein AutoDEGREE outperforms BERT_QA significantly across all test suites.

from ACE to GENEVA. Note that this overlap is not exhaustively complete. Furthermore, the mapping can be many-to-one and one-to-many in nature.

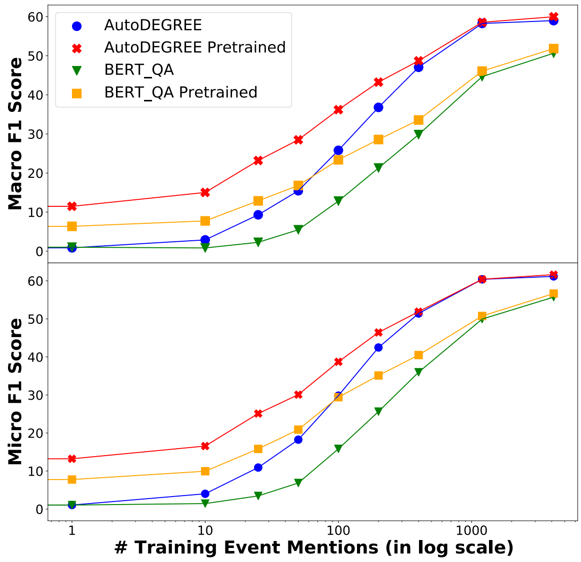

Figure 10: Model performance in macro F1 (top) and micro F1 (bottom) scores against the number of training event mentions (log-scale) for the low resource suite. Here we majorly compare the impact of pre-training on the model performance.

G Impact of Pre-training

In this section, we explore the impact of pre-training our models on previous datasets like ACE/ERE and evaluating them on the GENEVA benchmarking setups. Currently, we only report the model performance for our proposed model AutoDEGREE and a classification baseline model of BERT_QA.[12] Figures 10 and 11 show the impact of pre-training on the low-resource and few-shot test suites respectively.

We observe that pre-training helps model performance by 5-10 F1 points, and naturally in the low-data regime. But the gains diminish and are almost negligible when the number of training event mentions increases. Also, the zero-shot performance for the pretrained models isn't as impressive with AutoDEGREE achieving a micro F1 of 12.83 and BERT_QA achieving a score of 6.82 respectively, despite being fully trained on the ACE dataset. Poor zero-shot performance and diminishing performance gains indicate that GENEVA is distributionally distinct from the ACE dataset, which makes it challenging to achieve good model performance on GENEVA merely via transfer learning.

[12] We use BERT-Base as the PLM for these experiments.

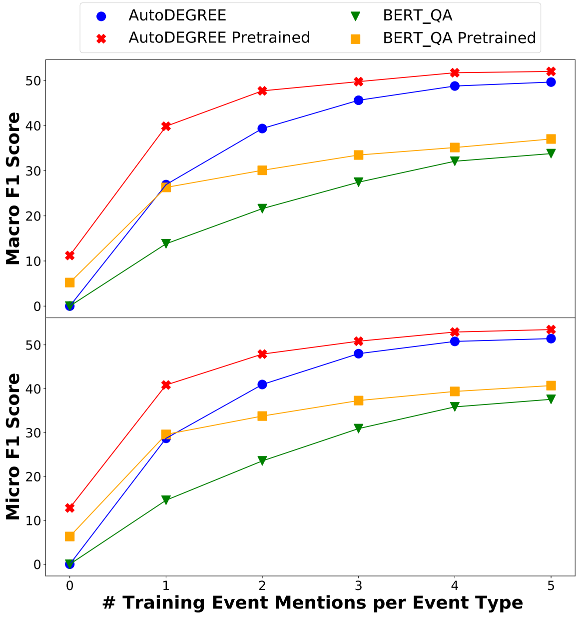

Figure 11: Model performance in macro F1 (top) and micro F1 (bottom) scores against the number of training event mentions per event for the few-shot test suite. Here we majorly compare the impact of pre-training on the model performance.

H Human Validation Experiment Setup

We present the human annotators with three sentences - one primary and two candidates - and ask them if the event type described in the primary sentence is similar to the event types in either of the candidates or distinct from both (Example in Appendix H). One candidate is chosen as a sentence from one of the frames merged with the primary event, while the other candidate is chosen from a similar unmerged frame, which is a sibling event of the primary event discovered from the event ontology. The annotator chooses between three options - Candidate 1, Candidate 2, or None. We provide an example of the annotation setup used for the human validation experiment conducted as part of GENEVA creation process in Table 8.

I Hyperparameters and Experimental Setup

In this section, we provide details about the experimental setups and training details for various EAE models we mentioned in our work.

I.1 AutoDEGREE

We closely follow the training setup by DEGREE for training the AutoDEGREE models. We run experiments for AutoDEGREE on a NVIDIA GeForce RTX 2080 Ti machine with support for 8 GPUs. We present the complete range of hyperparameter details in Table 9. We deploy early stopping criteria for stopping the model training.

I.2 BERT_QA

We mostly follow the original experimental setup and hyperparameters as described in Du and Cardie (2020). We use BERT-LARGE instead of the original BERT-BASE to ensure that the PLMs are of comparable sizes for AutoDEGREE and BERT_QA. We run experiments for this model on a NVIDIA A100-SXM4-40GB machine with support for 4 GPUs. A more comprensive list of hyperparameters is provided in Table 10.

I.3 TANL

We report the hyperparameter settings for the TANL experiments in Table 11. We make optimization changes in the provided source code of TANL to include multiple triggers in a single sentence. Experiments for TANL were run on a NVIDIA GeForce RTX 2080 Ti machine with support for 8 GPUs.

I.4 DyGIE++

We report the hyperparameter settings for the DyGIE++ experiments in Table 12. Experiments for DyGIE++ were run on a NVIDIA GeForce RTX 2080 Ti machine with support for 4 GPUs.

I.5 OneIE

We report the hyperparameter settings for the OneIE experiments in Table 13. Experiments for OneIE were run on a NVIDIA GeForce RTX 2080 Ti machine with support for 4 GPUs.

J Complete Results

In this section, we present the exhaustive set of results for each of the runs for the different benchmarking suites. We show the results for the low resource and few-shot setting are shown in Tables 14 and 15 respectively. Tables 16 and 17 display the results for the zero-shot and cross-type transfer settings respectively.

	Sentence	Event Trigger
Primary	Both villages offer good waterfront restaurants with homestyle Chinese food, principally seafood fresh from the tank.	offer
Candidate 1	It gives an overview of Macau's history and its daily life and traditions.	gives
Candidate 2	He should do more to reduce tax rates on wealth and income, in recognition of the fact that those cuts yield higher, not lower, revenues.	revenues

Table 8: Illustration of the human validation setup for one annotation. This setup is used for evaluating the merging operation done in the creation of GENEVA.

PLM	BART-Large
Training Batch Size	6
Eval Batch Size	12
Learning Rate	1×10^{-5}
Weight Decay	1×10^{-5}
# Warmup Epochs	5
Gradient Clipping	5
Max Training Epochs	50
# Accumulation Steps	1
Beam Size	1
Max Sequence Length	400
Max Output Length	50

Table 9: Hyperparameter details for AutoDEGREE model.

PLM	BERT-Large
Training Batch Size	24
Eval Batch Size	16
Learning Rate	1×10^{-5}
# Training Epochs	8^*
# Evaluations per Epoch	5
Max Sequence Length	300
Max Answer Length	30
N-Best Size	20

Table 10: Hyperparameter details for BERT_QA model. $*$ indicates that we increase the training epochs upto 25 as we reduce the training data for low-resource and few-shot settings.

PLM	T5-Base
Training Batch Size	8
Eval Batch Size	12
Learning Rate	5×10^{-4}
# Training Epochs	4^*
Evaluation per # Steps	100
Max Sequence Length	256
# Beams	8

Table 11: Hyperparameter details for TANL model. $*$ indicates that we increase the training epochs upto 100 as we reduce the training data for low-resource and few-shot settings.

PLM	BERT-Large
Training Batch Size	6
Eval Batch Size	12
Learning Rate	2×10^{-5}
# Training Epochs	200^*
Evaluation per # Epoch	1
Max Sequence Length	175
# Beams	8

Table 12: Hyperparameter details for DyGIE++ model. $*$ indicates that we increase the training epochs upto 200 as we reduce the training data for low-resource and few-shot settings.

PLM	BERT-Large
Training Batch Size	6
Eval Batch Size	12
Learning Rate	1×10^{-5}
# Training Epochs	150^{*}
Evaluation per # Epoch	1
Max Sequence Length	175
# Beams	8

Table 13: Hyperparameter details for OneIE model. $*$ indicates that we increase the training epochs upto 150 as we reduce the training data for low-resource and few-shot settings.

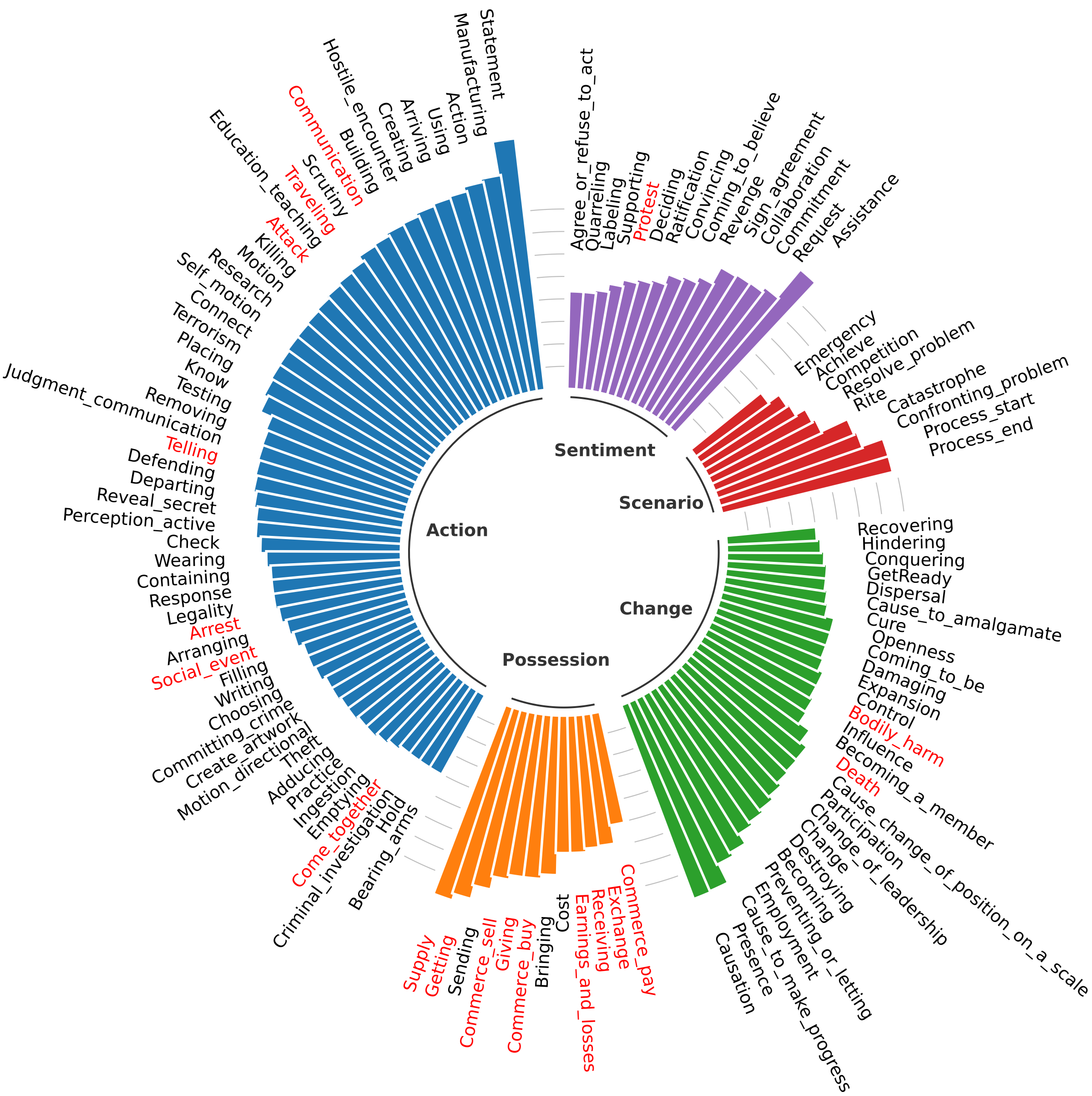

Figure 12: Circular bar plot for the various event types present in the GENEVAdataset organized into abstract event types. The height of each bar is proportional to the number of event mentions for that event (height is in log-scale). Bar labels colored in red are the set of overlapping event types mapped from the ACE dataset.

# Training Event Mentions		AutoDEGREE					BERT_QA				
1	Micro	0.07	0.00	0.15	5.09	0.00	0.75	0.63	0.51	0.00	3.40
	Macro	0.29	0.00	0.22	3.81	0.00	1.00	0.70	0.81	0.00	2.57
10	Micro	1.33	3.99	6.56	1.75	6.45	0.37	2.34	1.30	1.95	1.27
	Macro	1.03	2.78	4.77	0.74	5.14	0.16	1.60	0.79	0.98	0.61
25	Micro	8.69	11.32	12.44	16.51	5.64	2.26	4.39	4.79	4.79	1.07
	Macro	6.67	8.84	10.59	14.87	5.56	1.34	2.69	3.00	3.36	0.97
50	Micro	18.28	21.75	15.78	19.48	15.97	6.85	6.72	7.61	6.55	6.59
	Macro	15.49	17.16	14.14	17.28	13.26	5.51	4.90	6.42	5.51	5.05
100	Micro	32.51	33.16	30.37	27.84	25.25	18.00	15.52	14.15	15.10	16.40
	Macro	29.31	29.95	23.90	23.41	22.47	16.02	13.05	10.24	11.82	12.96
200	Micro	45.21	40.31	41.38	45.21	40.31	26.36	27.01	22.03	26.07	26.66
	Macro	38.72	35.31	35.96	38.72	35.31	20.94	23.43	19.07	20.55	22.46
400	Micro	50.00	52.25	51.39	51.42	52.06	37.28	37.61	36.91	35.65	32.40
	Macro	45.15	47.83	47.03	46.79	48.52	31.04	30.99	30.79	29.67	26.68
1200	Micro	61.16	59.35	60.25	60.64	60.60	47.68	52.93	49.01	48.90	51.24
	Macro	58.71	56.45	58.10	58.89	59.21	42.19	47.17	44.65	42.25	47.10
4132	Micro	61.35	61.20	61.20	60.92	61.16	55.43	56.94	55.66	54.40	56.15
	Macro	58.76	59.18	59.18	58.28	59.60	50.20	52.02	50.54	49.69	50.86
		DyGIE++					**TANL**				
1	Micro	0.01	0.15	0.00	0.73	0.57	0.07	0.22	0.20	0.97	1.52
	Macro	0.01	0.08	0.00	0.19	0.51	0.29	0.08	0.07	0.70	1.16
10	Micro	0.00	0.00	0.00	0.00	0.00	0.47	1.03	7.06	1.38	4.55
	Macro	0.00	0.00	0.00	0.00	0.00	0.52	0.72	2.54	1.42	2.52
25	Micro	0.52	0.15	0.37	1.98	0.01	6.77	8.98	8.34	13.26	4.65
	Macro	0.36	0.07	0.33	1.99	0.02	3.92	4.36	4.82	8.38	4.15
50	Micro	1.62	1.83	1.18	0.52	0.96	12.36	16.81	14.30	18.49	13.14
	Macro	1.56	1.77	1.40	0.49	0.73	6.76	9.35	9.78	10.00	8.11
100	Micro	6.24	6.28	7.46	4.94	4.38	27.44	24.09	28.50	26.05	25.44
	Macro	4.12	4.52	4.12	3.78	4.29	17.08	14.31	15.68	16.37	16.28
200	Micro	16.17	13.99	12.81	15.17	12.06	40.86	41.19	36.94	41.77	39.10
	Macro	9.62	10.18	8.50	9.01	6.62	27.01	28.99	25.61	26.08	25.25
400	Micro	28.44	29.42	32.75	29.42	29.61	49.84	50.48	50.77	50.44	51.01
	Macro	17.95	21.20	21.40	19.75	19.30	35.76	35.36	36.86	35.85	36.01
1200	Micro	57.00	56.49	55.29	58.24	57.40	63.97	61.69	59.98	60.04	61.79
	Macro	46.52	44.80	45.02	46.13	46.85	51.46	47.92	45.85	45.30	47.44
4132	Micro	66.07	67.27	66.42	66.58	66.77	68.78	68.94	68.18	69.07	68.17
	Macro	54.88	57.00	55.35	55.51	55.23	58.67	57.90	58.20	58.31	58.93

Table 14: Complete set of results of the 5 different runs for all models for the low resource test suite. Here Micro is the micro F1 score and Macro is the macro F1 score.

# Training Event Mentions per Event Type		AutoDEGREE					BERT_QA				
1	Micro	30.75	31.31	28.49	31.46	21.42	15.98	15.09	11.97	14.16	15.75
	Macro	28.62	29.64	27.95	29.73	18.67	16.38	13.48	11.00	13.21	14.97
2	Micro	40.51	39.16	40.49	40.89	43.75	26.42	22.79	27.15	21.42	19.97
	Macro	39.39	39.17	38.62	38.37	41.20	23.38	20.98	24.72	20.06	18.84
3	Micro	48.75	47.19	47.25	49.61	47.16	31.28	31.69	28.62	31.06	31.88
	Macro	46.19	44.92	44.88	46.98	45.06	28.31	27.66	26.28	27.06	28.00
4	Micro	51.93	50.48	50.57	50.56	50.37	36.70	36.47	33.53	36.31	36.27
	Macro	49.68	48.00	48.80	47.75	49.64	32.22	32.97	30.45	31.64	33.20
5	Micro	51.56	49.67	51.98	51.91	51.97	34.39	37.09	39.12	37.36	39.93
	Macro	50.98	48.16	49.96	49.69	49.42	30.88	33.88	35.84	32.75	35.60
		DyGIE++					**TANL**				
1	Micro	2.03	1.54	1.98	1.97	3.58	20.50	22.53	17.88	21.10	19.11
	Macro	2.79	2.13	2.48	2.33	3.82	15.87	19.14	14.80	17.75	15.19
2	Micro	5.71	4.15	5.75	4.44	6.32	33.12	33.59	36.28	33.18	36.27
	Macro	6.24	5.42	6.22	5.18	8.36	29.00	28.01	31.47	29.19	31.74
3	Micro	10.33	11.27	10.20	13.90	9.13	40.36	42.91	39.30	45.95	43.08
	Macro	11.56	12.80	11.75	14.59	10.23	36.52	38.18	34.55	40.93	37.27
4	Micro	14.50	17.21	11.93	13.51	11.25	43.55	45.95	45.27	47.56	47.35
	Macro	15.63	16.83	13.44	14.30	13.59	40.62	40.80	42.54	43.30	44.01
5	Micro	14.69	16.33	18.90	17.56	21.77	48.97	50.43	49.04	50.51	51.44
	Macro	16.82	17.48	19.46	19.75	22.27	44.09	46.20	44.87	45.18	47.65

Table 15: Complete set of results of the 5 different runs for all models for the few shot test suite. Here Micro is the micro F1 score and Macro is the macro F1 score.

# Training Events		AutoDEGREE					BERT_QA				
1	Micro	14.87	13.99	14.10	14.12	12.46	5.44	4.37	5.63	4.83	5.76
	Macro	14.48	13.38	12.77	12.01	10.43	3.55	2.82	2.99	3.16	3.06
5	Micro	33.68	31.56	33.32	32.62	34.11	24.92	23.69	22.11	23.52	21.51
	Macro	33.23	30.72	33.41	30.92	33.18	23.88	20.90	18.18	19.86	17.15
10	Micro	36.79	34.72	36.90	33.64	35.31	23.30	23.48	22.68	23.45	23.25
	Macro	36.43	33.00	36.19	34.10	34.30	20.20	20.05	19.33	20.61	19.47

Table 16: Complete set of results of the 5 different runs for all models for the zero-shot test suite. Here Micro is the micro F1 score and Macro is the macro F1 score.

	AutoDEGREE					BERT_QA				
Micro	28.28	25.58	27.05	28.73	26.67	8.19	4.44	10.69	7.24	8.58
Macro	28.51	26.23	25.58	28.98	27.03	8.97	3.35	10.76	7.24	9.88

Table 17: Complete set of results of the 5 different runs for all models for the cross-type transfer test suite. Here Micro is the micro F1 score and Macro is the macro F1 score.