GENEVA: Benchmarking Generalizability for Event Argument Extraction with Hundreds of Event Types and Argument Roles

Tanmay Parekh[†] **I-Hung Hsu**[‡] **Kuan-Hao Huang**[†]
Kai-Wei Chang[†] **Nanyun Peng**[†]
[†]Computer Science Department, University of California, Los Angeles
[‡]Information Science Institute, University of Southern California
{tparekh, khhuang, kwchang, violetpeng}@cs.ucla.edu
{ihunghsu}@isi.edu

Abstract

Recent works in Event Argument Extraction (EAE) have focused on improving model generalizability to cater to new events and domains. However, standard benchmarking datasets like ACE and ERE cover less than 40 event types and 25 entity-centric argument roles. Limited diversity and coverage hinder these datasets from adequately evaluating the generalizability of EAE models. In this paper, we first contribute by creating a large and diverse EAE ontology. This ontology is created by transforming FrameNet, a comprehensive semantic role labeling (SRL) dataset for EAE, by exploiting the similarity between these two tasks. Then, exhaustive human expert annotations are collected to build the ontology, concluding with 115 events and 220 argument roles, with a significant portion of roles not being entities. We utilize this ontology to further introduce GENEVA, a diverse generalizability benchmarking dataset comprising four test suites, aimed at evaluating models' ability to handle limited data and unseen event type generalization. We benchmark six EAE models from various families. The results show that owing to non-entity argument roles, even the best-performing model can only achieve 39% F1 score, indicating how GENEVA provides new challenges for generalization in EAE. Overall, our large and diverse EAE ontology can aid in creating more comprehensive future resources, while GENEVA is a challenging benchmarking dataset encouraging further research for improving generalizability in EAE. The code and data can be found at https://github.com/PlusLabNLP/GENEVA.

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 is a classic topic (Sundheim, 1992) and elemental for a wide range of applications like building knowledge graphs (Zhang et al., 2020),

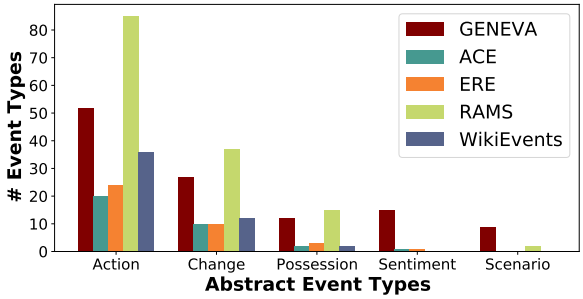

Figure 1: Distribution of event types into various abstract event types[1] for GENEVA, ACE, ERE, RAMS, and WikiEvents datasets. We observe that GENEVA is relatively more diverse than the other datasets.

question answering (Berant et al., 2014), and others (Hogenboom et al., 2016; Wen et al., 2021; Yang et al., 2019b). Recent works have focused on building generalizable EAE models (Huang et al., 2018; Lyu et al., 2021; Sainz et al., 2022) and they utilize existing datasets like ACE (Doddington et al., 2004) and ERE (Song et al., 2015) for benchmarking. However, as shown in Figure 1, these datasets have limited diversity as they focus only on two abstract types,[1] Action and Change. Furthermore, they have restricted coverage as they only comprise argument roles that are entities. The limited diversity and coverage restrict the ability of these existing datasets to robustly evaluate the generalizability of EAE models. Toward this end, we propose a new generalizability benchmarking dataset in our work.

To build a strong comprehensive benchmarking dataset, we first create a large and diverse ontology. Creating such an ontology from scratch is time-consuming and requires expert knowledge. To reduce human effort, we exploit the shared properties between semantic role labeling (SRL) and EAE (Aguilar et al., 2014) and leverage a diverse and exhaustive SRL dataset, FrameNet (Baker et al., 1998), to build the ontology. Through extensive human expert annotations, we design mappings

[1]Abstract event types are defined as the top nodes of the event ontology created by MAVEN (Wang et al., 2020).

that transform the FrameNet schema to a large and diverse EAE ontology, spanning 115 event types from five different abstract types. Our ontology is also comprehensive, comprising 220 argument roles with a significant 37% of roles as non-entities.

Utilizing this ontology, we create GENEVA - a **G**eneralizability B**EN**chmarking Dataset for **EV**ent **A**rgument Extraction. We exploit the human-curated ontology mappings to transfer FrameNet data for EAE to build GENEVA. We further perform several human validation assessments to ensure high annotation quality. GENEVA comprises four test suites to assess the models' ability to learn from limited training data and generalize to unseen event types. These test suites are distinctly different based on the training and test data creation – (1) low resource, (2) few-shot, (3) zero-shot, and (4) cross-type transfer settings.

We use these test suites to benchmark various classes of EAE models - traditional classification-based models (Wadden et al., 2019; Lin et al., 2020; Wang et al., 2022a), question-answering-based models (Du and Cardie, 2020), and generative approaches (Paolini et al., 2021; Hsu et al., 2022b). We also introduce new automated refinements in the low resource state-of-the-art model DEGREE (Hsu et al., 2022b) to generalize and scale up its manual input prompts. Experiments reveal that DEGREE performs the best and exhibits the best generalizability. However, owing to non-entity arguments in GENEVA, DEGREE achieves an F1 score of only 39% on the zero-shot suite. Under a similar setup on ACE, DEGREE achieves 53%, indicating how GENEVA poses additional challenges for generalizability benchmarking.

To summarize, we make the following contributions. We construct a diverse and comprehensive EAE ontology introducing non-entity argument roles. This ontology can be utilized further to develop more comprehensive datasets for EAE. In addition, we propose a generalizability evaluation dataset GENEVA and benchmark various recent EAE models. Finally, we show how GENEVA is a challenging dataset, thus, encouraging future research for generalization in EAE.

2 Related Work

Event Extraction Datasets and Ontologies: The earliest datasets in event extraction date back to MUC (Sundheim, 1992; Grishman and Sundheim, 1996). Doddington et al. (2004) introduced the standard dataset ACE while restricting the ontology to focus on entity-centric arguments. The ACE ontology was further simplified and extended to ERE (Song et al., 2015) and various TAC KBP Challenges (Ellis et al., 2014, 2015; Getman et al., 2017). These datasets cover a small and restricted set of event types and argument roles with limited diversity. Later, MAVEN (Wang et al., 2020) introduced a massive dataset spanning a wide range of event types. However, its ontology is limited to the task of Event Detection[2] and does not contain argument roles. Recent works have introduced document-level EAE datasets like RAMS (Ebner et al., 2020), WikiEvents (Li et al., 2021), and DocEE (Tong et al., 2022); but their ontologies are also entity-centric, and their event coverage is limited to specific abstract event types (Figure 1). In our work, we focus on building a diverse and comprehensive dataset for benchmarking generalizability for sentence-level EAE.

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, these approaches exhibit poor generalizability in the low-data setting (Liu et al., 2020; Hsu et al., 2022b). 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 structured prediction to boost generalizability (Schick and Schütze, 2021a,b; Paolini et al., 2021; Li et al., 2021). Another set of works transfers knowledge from similar tasks like abstract meaning representation and semantic role labeling (Huang et al., 2018; Lyu et al., 2021; Zhang et al., 2021). DEGREE (Hsu et al., 2022b) is a recently introduced state-of-the-art generative model which has shown the best performance in the limited data regime. In our work, we benchmark the generalizability of various classes of old and new models on our dataset.

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

3 Ontology Creation

Event annotations start with ontology creation, which defines the scope of the events and their corresponding argument roles of interests. Towards this end, we aim to construct a large ontology of diverse event types with an exhaustive set of event argument roles. However, it is a challenging and tedious task that requires extensive expert supervision if building from scratch. To reduce human effort while maintaining high quality, we leverage the shared properties of SRL and EAE and utilize a diverse and comprehensive SRL dataset — FrameNet to design our ontology. We first re-iterate the EAE terminologies we follow (§ 3.1) and then describe how FrameNet aids our ontology design (§ 3.2). Finally, we present our steps for creating the final ontology in § 3.3 and ontology statistics in § 3.4.

3.1 Task Definition

We follow the definition of **event** as a class attribute with values such as *occurrence, state, or reporting* (Pustejovsky et al., 2003; Han et al., 2021). **Event Triggers** are word phrases that best express the occurrence of an event in a sentence. Following the early works of MUC (Sundheim, 1992; Grishman and Sundheim, 1996), **event arguments** are defined as participants in the event which provide specific and salient information about the event. **Event argument role** is the semantic category of the information the event argument provides. We provide an illustration in Figure 2 describing an event about *"Destroying"*, where the event trigger is *obliterated*, and the event consists of argument roles — *Cause* and *Patient*.

It is worth mentioning that these definitions are disparate from the ones that previous works like ACE, and its inheritors, ERE and RAMS, follow. In ACE, the scope of events is restricted to the attribute of occurrence only, and event arguments are restricted to entities, wherein **entities** are defined as objects in the world. For example, in Figure 2, *the subsequent explosions* isn't an entity and will not be considered an argument as per ACE definitions. Consequently, *Cause* won't be part of their ontology. This exclusion of non-entities leads to incomplete information extraction of the event. In our work, we follow MUC to consider a broader range of events and event arguments.

3.2 FrameNet for EAE

To overcome the challenge of constructing an event ontology from scratch, we aim to leverage

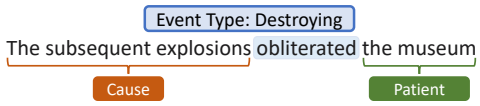

Figure 2: An illustration of EAE for the Destroying event comprising argument roles of Cause and Patient.

FrameNet, a semantic role labeling (SRL) dataset, to help our ontology creation. The similarity between SRL and EAE (Aguilar et al., 2014) provides us with the ground for leveraging FrameNet. SRL assigns semantic roles to phrases in the sentence, while EAE extracts event-specific arguments and their roles from the sentence. Hence, *selecting event-related parts* of a fine-grained annotated SRL dataset can be considered as an exhaustively annotated resource for EAE.

We choose FrameNet[3] (Baker et al., 1998) as the auxiliary SRL dataset since it is one of the most comprehensive SRL resources. It comprises 1200+ semantic frames (Fillmore et al., 1976), where a **frame** is a holistic background that unites similar words. Each frame is composed of frame-specific semantic roles (**frame elements**) and is evoked by specific sets of words (**lexical units**).

To transfer FrameNet's schema into an EAE ontology, we map *frames* as events, *lexical units* as event triggers, and *frame elements* as argument roles. However, this basic mapping is inaccurate and has shortcomings since *not all frames are events*, and *not all frame elements are argument roles* per the definitions in § 3.1. We highlight these shortcomings in Figure 3, which enlists some FrameNet frames and frame elements for the *Arrest* frame. Based on EAE definitions, only some frames like *Arrest, Travel, etc* (highlighted in yellow) can be mapped as events, and similarly, limited frame elements like *Authorities, Charges, etc* (highlighted in green) are mappable as argument roles.

3.3 Building the EAE Ontology

To overcome the shortcomings of the basic mapping, we follow a two-step approach (Figure 4). First, we build an event ontology for accurately mapping frames to events. Then, we augment this ontology with argument roles by building an event argument ontology. We describe these steps below.

Event Ontology: In order to build the event on-

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

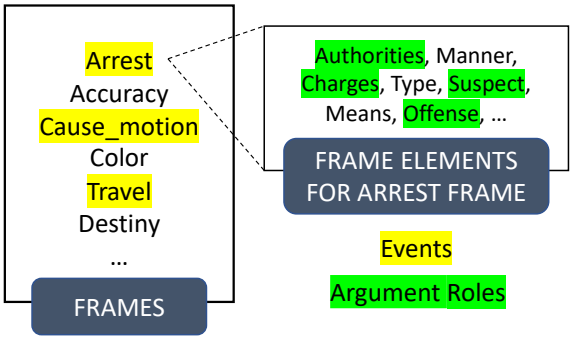

Figure 3: Illustration of challenges in using FrameNet for EAE - Not all frames are events and not all frame elements are argument roles.

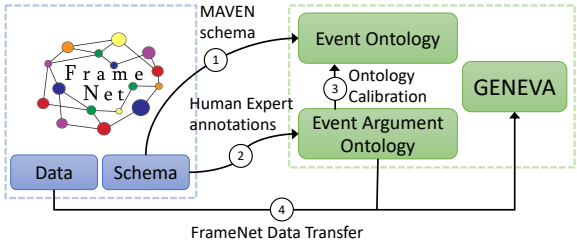

Figure 4: Illustration of the GENEVA creation from FrameNet labeled sequentially by the crucial steps.

tology, we utilize the event mapping designed by MAVEN (Wang et al., 2020), which is an event detection dataset. They first recursively filter frames having a relation with the "Event" frame in FrameNet. Then they manually filter and merge frames based on the definitions, resulting in an event ontology comprising 168 event types mapped from 289 filtered frames.

Event Argument Ontology: In order to augment argument roles to the event ontology, we perform an extensive human expert annotation process. The goal of this annotation process is to create an argument mapping from FrameNet to our ontology by filtering and merging frame elements. We describe this annotation process below.

Annotation Instructions: Annotators are provided with a list of frame elements along with their descriptions for each frame in the event ontology. [4] They are also provided with definitions for events and argument roles as discussed in Section 3.1. Based on these definitions, they are asked to annotate each frame element as (a) not argument role, (b) argument role, or (c) merge with existing argument role (and mention the argument role to merge with). To ensure arguments are salient, annotators are instructed to filter out frame elements that are super generic (e.g. Time, Place, Purpose) unless they are relevant to the event. Ambiguous cases are flagged and commonly reviewed at a later stage.

Additionally, annotators are asked to classify each argument role as an entity or not. This additional annotation provides flexibility for quick conversion of the ontology to ACE definitions. Figure 14 in the Appendix provides an illustration of these instructions and the annotation process.

Annotation Results: We recruit two human experts who are well-versed in the field of event extraction. We conduct three rounds of annotations and discussions to improve consistency and ensure a high inter-annotator agreement (IAA). The final IAA measured as Cohen's Kappa (McHugh, 2012) was 0.82 for mapping frame elements and 0.94 for entity classification. A total of 3,729 frame elements from 289 frames were examined as part of the annotation process. About 63% frame elements were filtered out, 14% were merged, and the remaining 23% constitute as argument roles.

Event Ontology Calibration: The MAVEN event ontology is created independent of the argument roles. This leads to some inaccuracies in their ontology wherein two frames with disparate sets of argument roles are mapped as a single event. For example, *Surrendering_possession* and *Surrendering* frames are merged together despite having different argument roles. Based on our human expert-curated event argument ontology, we rectify these inaccuracies (roughly 8% of the event ontology) and create our final ontology.

3.4 Ontology Statistics

We present the statistics of our full ontology in Table 1 and compare it with existing ACE (Doddington et al., 2004) and RAMS (Ebner et al., 2020) ontologies. But as we will specify in § 4.1, we use a subset of this ontology[5] for creating GENEVA. Hence, we also include the statistics of the GENEVA ontology in the last column in Table 1. Overall, our curated full ontology is the largest and most comprehensive as it comprises 179 event types and 362 argument roles. Defining *abstract event types* as the top nodes of the ontology tree created by MAVEN (Wang et al., 2020), we show that our ontology spans 5 different abstract types and is the most diverse. We organize our ontology into a hierarchy of these abstract

[4] Event ontology frames can be viewed as candidate events.

[5] We will release both full and GENEVA ontologies to facilitate future study.

	ACE	RAMS	Full	GENEVA
# Event Types	33	139	179	115
# Abstract Event Types	2	3	5	5
# Argument Roles (AR)	22	65	362	220
Avg. # AR per Event	4.75	3.76	4.82	3.97
% Entity AR	100%	100%	65%	63%
% Non-Entity AR	0%	0%	35%	37%

Table 1: Full and GENEVA ontology Statistics. AR = Argument Role. An ontology covers an abstract type if it has $5+$ events of that abstract type. Entity AR refers to argument roles that are entities.

event types in Appendix A.3. Our ontology is also dense with an average of 4.82 argument roles per event type. Finally, we note that a significant 35% of the event argument roles in our ontology are non-entities. This demonstrates how our ontology covers a broader and more comprehensive range of argument roles than other ontologies following ACE definitions of entity-centric argument roles.

4 GENEVA Dataset

Previous EAE datasets for evaluating generalizability like ACE and ERE have limited event diversity and are restricted to entity-centric arguments. To overcome these issues, we utilize our ontology to construct a new generalizability benchmarking dataset GENEVA comprising four specialized test suites. We describe our data creation process in § 4.1, provide data statistics in § 4.2 and discuss out test suites in § 4.3.

4.1 Creation of GENEVA

Since annotating EAE data for our large ontology is an expensive process, we leverage the annotated dataset of FrameNet to create GENEVA (Figure 4). We utilize the previously designed ontology mappings to repurpose the annotated sentences from FrameNet for EAE by mapping frames to corresponding events, lexical units to event triggers, and frame elements to corresponding arguments. Unmapped frames and frame elements (not in the ontology) are filtered out from the dataset. Since FrameNet doesn't provide annotations for all frames, some events from the full ontology are not present in our dataset (e.g. *Military_Operation*). Additionally, to aid better evaluation, we remove events that have less than 5 event mentions (e.g. *Lighting*). Finally, GENEVA comprises 115 event types and 220 argument roles. Some examples are provided in Figure 10 (Appendix).

Human Validation: We ensure the high quality of our dataset by conducting two human assessments: (1) *Ontology Quality Assessment*: We present the human annotators with three sentences - one primary and two candidates - and ask them if the event in the primary sentence is similar to the events in either of the candidates or distinct from both (Example in Appendix F). One candidate sentence is chosen from the frame merged with the primary event, while the other candidate is chosen from a similar unmerged sister frame. The annotators chose the merged frame candidates 87% of the times, demonstrating the high quality of the ontology mappings. This validation was done by three annotators over 61 triplets with 0.7 IAA measured by Fleiss' kappa (Fleiss, 1971).
(2) *Annotation Comprehensiveness Assessment*: Human annotators are presented with annotated samples from our dataset and they are asked to report if there are any arguments in the sentence that have not been annotated. The annotation is considered comprehensive if all arguments are annotated correctly. The annotators reported that the annotations were 89% comprehensive, ensuring high dataset quality. Corrections majorly comprise ambiguous cases and incorrect role labels. This assessment was done by two experts over 100 sampled annotations with 0.93 IAA (Cohen's kappa).

Dataset	#Event Types	#Arg Types	Avg. Event Mentions	Avg. Arg Mentions
ACE	33	22	153.18	274.55
ERE	38	21	191.76	499
GENEVA	115	220	65.26	55.77

Table 2: Statistics for different EAE datasets for benchmarking generalizability. The second and third columns are the unique number of event types and argument roles. The last two columns indicate the average number of mentions per event and argument role.

4.2 Data Analysis

Overall, GENEVA is a dense, challenging, and diverse EAE dataset with good coverage. These characteristics make GENEVA better-suited than existing datasets like ACE/ERE for evaluating the generalizability of EAE models. The major statistics for GENEVA are shown in Table 2 along with its comparison with ACE and ERE. We provide more discussions about the characteristics of our dataset as follows.

Diverse: GENEVA has wide coverage with a tripled number of event types and 10 times the number of argument roles relative to ACE/ERE.

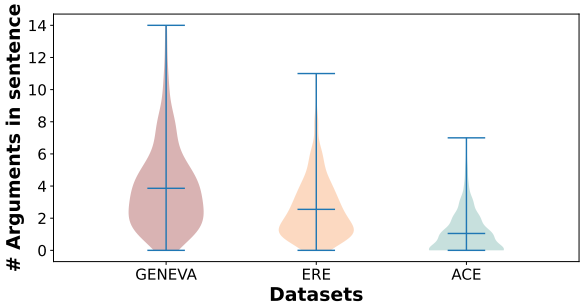

Figure 5: Violin plots for number of arguments per sentence for ACE, ERE and GENEVA datasets.

Figure 1 further depicts how ACE/ERE focus only on specific abstractions Action and Change, while GENEVA is the most diverse with events ranging from 5 abstract types.

Challenging: The average number of mentions per event type and argument role (Table 2) is relatively less for GENEVA. Consequently, EAE models need to train from fewer examples on average which makes training more challenging.

Dense: We plot the distribution of arguments per sentence[6] for ACE, ERE, and GENEVA in Figure 5. We note that GENEVA has the highest density of 4 argument mentions per sentence. Both ACE and ERE have more than 70% sentences with up to 2 arguments. In contrast, GENEVA is denser with almost 50% sentences having 3 or more arguments.

Coverage: Qualitatively, we show some coverage of diverse examples in Figure 9 (Appendix) and provide coverage for all events categorized by their abstraction in Figure 14 (Appendix). We observe frequent events like Statement, Arriving, Action while Recovering, Emergency, Hindering are less-frequent events. In terms of diversity of data sources, our data comprises a mixture of news articles, Wall Street Journal articles, books, Wikipedia, and other miscellaneous sources too.

4.3 Benchmarking Test Suites

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

Limited Training Data: This setting mimics the realistic scenario when there are fewer annotations available for the target events and evaluates the 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.[7] We record the model performance across a spectrum from extremely low resource ($n = 10$) to moderate 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 ($n = 5$).

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 training data comprises the top m events with most data, where m varies from 1 to 10.[8] The remaining 105 events are used for evaluation.
- Cross-type Transfer (CTT): We curate a training dataset comprising of events of a single abstraction category (e.g. Scenario), while the test dataset comprises events of all other abstraction types. This test suite also assesses models' transfer learning strength.

Data statistics for these suites are presented in Appendix A.2. For each setup, we sample 5 different datasets[9] and report the average model performance to account for the sampling variation.

5 Experimental Setup

We evaluate the generalizability of various EAE models on GENEVA. We describe these models in § 5.1 and the evaluation metrics in § 5.2.

5.1 Benchmarked Models

Overall, we benchmark six EAE models from various representative families are described below. Implementation details are specified in Appendix G.

Classification-based models: These traditional works predict arguments by learning to trace the argument span using a classification objective. We experiment with three models: (1) **DyGIE++** (Wadden et al., 2019), a traditional model utilizing multi-sentence BERT encodings and span graph propagation. (2) **OneIE** (Lin et al., 2020), a multi-tasking

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

[7] To discount the variance of the number of event mentions per sentence, we create the sampled training data such that each of them has a fixed number of n event mentions.

[8] We sample a fixed 450 sentences for training to remove the variance of dataset size for different m.

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

objective-based model exploiting global features for optimization. (3) **Query&Extract** (Wang et al., 2022a) utilizing the attention mechanism to extract arguments from argument role queries.

Question-Answering models: Several works formulate event extraction as a machine reading comprehension task. We consider one such model - (4) **BERT_QA** (Du and Cardie, 2020), a BERT-based model leveraging label semantics using a question-answering objective. In order to scale BERT_QA to the wide range of argument roles, we generate question queries of the form "*What is {arg-name}?*" for each argument role *{arg-name}*. (5) **TE** (Lyu et al., 2021), a zero-shot transfer model that utilizes an existing pre-trained textual entailment model to automatically extract events. Similar to BERT_QA, we design hypothesis questions as "*What is {arg-name}?*" for each argument role *{arg-name}*.

Generation-based models: Inspired by great strides in natural language generation, recent works frame EAE as a generation task using a language-modeling objective. We consider two such models: (6) **TANL** (Paolini et al., 2021), a multi-task language generation model which treats EAE as a translation task. (7) **DEGREE** (Hsu et al., 2022b), an encoder-decoder framework that extracts event arguments using natural language input prompts.

Automating DEGREE: DEGREE requires human effort for manually creating natural language prompts and thus, can not be directly deployed for the large set of event types in GENEVA. In our work, we undertake efforts to scale up DEGREE by proposing a set of automated refinements. The first refinement automates the event type description as "*The event type is {event-type}*" where *{event-type}* is the input event type. The second refinement automates the event template generation by splitting each argument into a separate self-referencing mini-template "*The {arg-name} is some {arg-name}*" where *{arg-name}* is the argument role. The final event-agnostic template is a simple concatenation of these mini-templates. We provide an illustration and ablation of these automated refinements for DEGREE in Appendix B.

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 a 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.

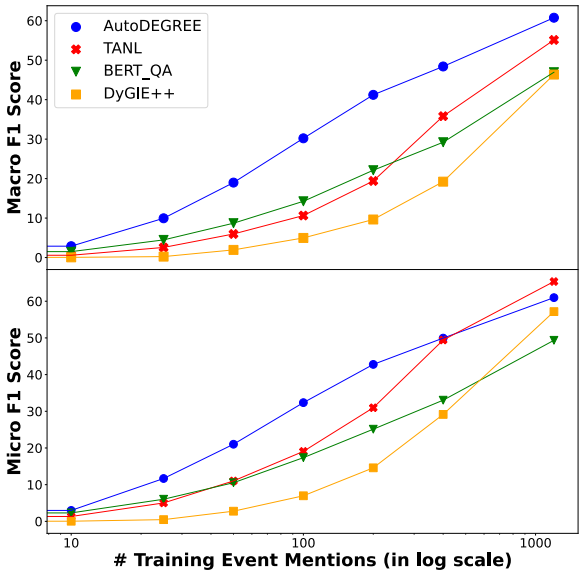

Figure 6: Model performance in macro F1 (top) and micro F1 (bottom) scores against the number of training event mentions (log-scale) for the low resource suite. Each datapoint is an average of 5 runs.

6 Results

Following § 4.3, 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 and Query&Extract models achieve poor micro F1 scores of just 30.03 and 40.41 while all other models achieve F1 scores above 55. This can be attributed to the inability of their model designs to effectively handle overlapping arguments.[10] Due to their inferior performance, we do not include OneIE and Query&Extract in the benchmarking results. We present the full results in Appendix H.

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 and micro F1 scores for the low resource test suite in Figure 6 and for the few-shot test suite in Figure 7 respectively. We observe that DEGREE outperforms all other models for both the test suites and shows superior generalizability. In general, we observe that generation-based models show better

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

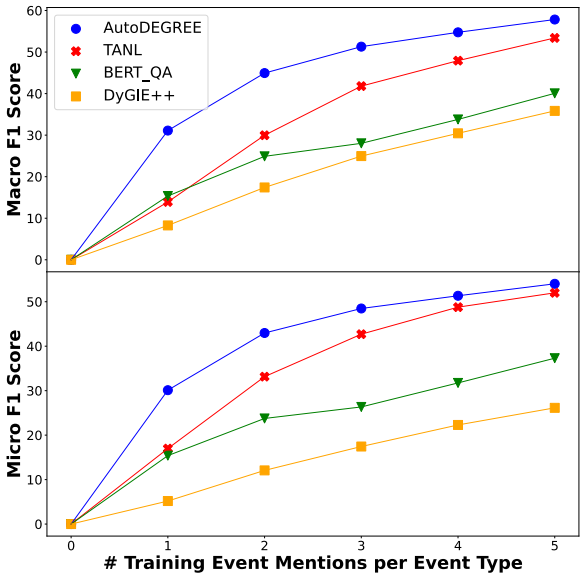

Figure 7: Model performance in macro F1 (top) and micro F1 (bottom) scores against the number of training event mentions per event for the few-shot suite. Each datapoint is an average of 5 runs.

generalization while on the other hand, traditional classification-based approaches show poor generalizability. This underlines the importance of using label semantics for better generalizability. We also detect a stark drop from micro to macro F1 scores for TANL and DyGIE++ in the low resource test suite. This indicates that these models are more easily biased toward high data events and do not generalize well uniformly across all events.

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. TE cannot be trained on additional EAE data, and hence we only report the pure zero-shot performance of this model.

From Table 3, we observe that DEGREE achieves the best scores across both test suites outperforming BERT_QA by a significant margin of almost 13-15% F1 points. Although TE is not comparable as it's a pure zero-shot model (without training on any data), it's performance is relatively super low in both settings. Thus, DEGREE shows superior transferability to unseen event types and argument roles.

Model	ZS-1	ZS-5	ZS-10	CTT
TE*	7.54	7.54	7.54	6.39
BERT_QA	5.05	21.53	24.24	11.17
DEGREE	**24.06**	**34.68**	**39.43**	**27.9**

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. Each datapoint is an average of 5 runs. *Not directly comparable as TE doesn't train on any data.

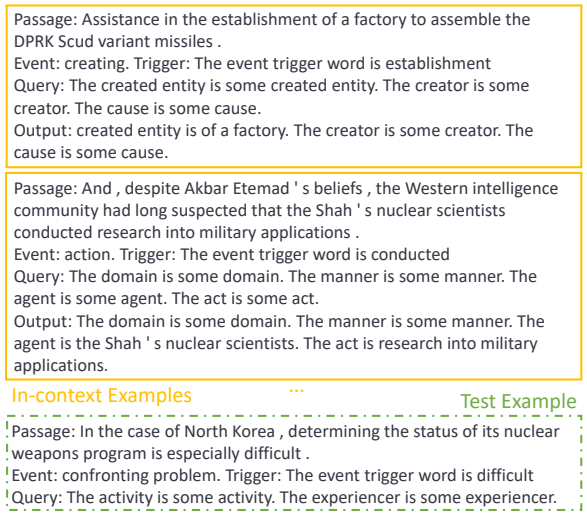

Figure 8: Illustration of the prompt used for evaluating GPT3.5-turbo. We provide 5 in-context examples (solid yellow) and provide the test example (dashed green).

7 Analysis

In this section, we provide analyses highlighting the various new challenges introduced by GENEVA. We discuss the performance of large language models, the introduction of non-entity argument roles, and model performance including Time and Place argument roles.

7.1 Large Language Model Performance

Recently, there has been an advent of Generative AI in the form of Large Language Models (LLMs) like GPT-3 (Brown et al., 2020), GPT-4, PaLM (Chowdhery et al., 2022), Code4Struct (Wang et al., 2022b), and many more. We evaluate one of these models GPT3.5-turbo on the task of EAE on the zero-shot test suite of GENEVA[11]. More specifically, we provide 5 in-context examples from top-10 events and evaluate test examples from the remaining 105 events. Our GPT-prompt template follows the DEGREE template wherein model re-

[11] Since we can't fine-tune LLMs on known event types, this is not the most fair comparison, but the closest one possible.

	LR-400		ZS-10	
	GENEVA	ACE	GENEVA	ACE
BERT_QA	33	-	24.2	46.7*
DEGREE	49.9	57.3*	39.4	53.3*

Table 4: Model performance in micro F1 for BERT_QA and DEGREE for low resource with 400 training mentions (LR-400) and zero-shot with 10 training events (ZS-10) test suites across GENEVA and ACE. *Reported from Hsu et al. (2022a).

places placeholders with arguments if present, else copies the original teample. An illustration is provided in Figure 8.

Despite the strong generation capability, GPT3.5-turbo achieves a mere **22.73** F1 score while DEGREE achieves **24.06** and **39.43** F1 scores in the ZS-1 and ZS-10 test suites respectively. Although these scores aren't directly comparable, it shows how GENEVA is quite challenging for LLMs in the zero-shot/few-shot setting.

7.2 New Challenge of Non-entity Roles

In Table 4, we show the model performances of BERT_QA and DEGREE on GENEVA and ACE under similar benchmarking setups. We note how both models exhibit relatively poor performance on GENEVA (especially the zero-shot test suite). To investigate this phenomenon, we break down the model performance based on entity and non-entity argument roles and show this analysis in Table 5. This ablation reveals a stark drop of 10-14% F1 points across all models when predicting non-entity arguments relative to entity-based arguments. This trend is observed consistently across all different test suites as well. We can attribute this difference in model performance to non-entity arguments being more abstract and having longer spans, in turn, being more challenging to predict accurately. Thus, owing to a significant 37% non-entity argument roles, GENEVA poses a new and interesting challenge for generalization in EAE.

7.3 GENEVA with Time and Place

In the original GENEVA dataset, we filtered super generic argument roles, but some of these roles like Time and Place are key for several downstream tasks. We include Time and Place arguments in GENEVA[12] and provide results of the models on the full dataset in Table 6. Compared to original GENEVA results in the same setting, we observe a slight dip in the model performance owing to the addition of extra arguments. Overall, the trend is similar where TANL performs the best and we observe better generalization in terms of macro F1 performance.

[12]We release this data for future development

	Entity	Non-entity	Δ
DEGREE	54.46	39.89	14.57
TANL	52.54	42.4	10.14
BERT_QA	36.71	24.86	11.85

Table 5: Breakdown of micro F1 scores into the entity and non-entity arguments for DEGREE, TANL, and BERT_QA models on the low resource setting with 400 training mentions. Δ denotes the difference.

Model	Micro F1	Macro F1
BERT_QA	52.97	50.16
DyGIE++	65.03	54.85
TANL	**71.17**	**65.18**
DEGREE	59.74	59.20

Table 6: Model performance in micro F1 and macro F1 scores for the full GENEVA dataset with Time and Place arguments.

7.4 Discussion

Overall, our generalizability benchmarking reveals various insights. First, generation-based models like DEGREE exhibit strong generalizability and establish a benchmark on our dataset. Second, macro score evaluation reveals how models like TANL and DyGIE++ can be easily biased toward high-data events. Finally, we show how GENEVA poses a new challenge in the form of non-entity arguments, encouraging further research for improving generalization in EAE.

8 Conclusion and Future Work

In our work, we exploit the shared relations between SRL and EAE to create a new large and diverse event argument ontology spanning 115 event types and 220 argument roles. This vast ontology can be used to create larger and more comprehensive resources for event extraction. We utilize this ontology to build a new generalizability benchmarking dataset GENEVA comprising four distinct test suites and benchmark EAE models from various families. Our results inspire further research of generative models for EAE to improve generalization. Finally, we show that GENEVA poses new challenges and anticipate future generalizability benchmarking efforts on our dataset.

Acknowledgements

We would like to thank Hritik Bansal, Di Wu, Sidi Lu, Derek Ma, Anh Mac, and Zhiyu Xie for their valuable insights, experimental setups, paper reviews, and constructive comments. We thank the anonymous reviewers for their feedback. This work was partially supported by NSF 2200274, AFOSR MURI via Grant #FA9550- 22-1-0380, Defense Advanced Research Project Agency (DARPA) grant #HR00112290103/HR0011260656, and a Cisco Sponsored Research Award.

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. Second, GENEVA is derived from an existing dataset FrameNet. Despite human validation efforts, there is no guarantee that all possible events in the sentence are exhaustively annotated.

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 training models for real-world applications.

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.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

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*.

Rujun Han, I-Hung Hsu, Jiao Sun, Julia Baylon, Qiang Ning, Dan Roth, and Nanyun Peng. 2021. ESTER: A machine reading comprehension dataset for reasoning about event semantic relations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*.

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. 2022a. DEGREE: A data-efficient generation-based event extraction model. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1890–1908, Seattle, United States. Association for Computational Linguistics.

I-Hung Hsu, Kuan-Hao Huang, Elizabeth Boschee, Scott Miller, Prem Natarajan, Kai-Wei Chang, and Nanyun Peng. 2022b. 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.

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.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

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)*.

James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, et al. 2003. The timebank corpus. In *Corpus linguistics*, volume 2003, page 40. Lancaster, UK.

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.

Sijia Wang, Mo Yu, Shiyu Chang, Lichao Sun, and Lifu Huang. 2022a. Query and extract: Refining event extraction as type-oriented binary decoding. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 169–182, Dublin, Ireland. 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.

Xingyao Wang, Sha Li, and Heng Ji. 2022b. Code4struct: Code generation for few-shot structured prediction from natural language. *arXiv preprint arXiv:2210.12810*.

Haoyang Wen, Ying Lin, Tuan Manh Lai, Xiaoman Pan, Sha Li, Xudong Lin, Ben Zhou, Manling Li, Haoyu Wang, Hongming Zhang, Xiaodong Yu, Alexander Dong, Zhenhailong Wang, Yi Ren Fung, Piyush Mishra, Qing Lyu, Dídac Surís, Brian Chen, Susan Windisch Brown, Martha Palmer, Chris Callison-Burch, Carl Vondrick, Jiawei Han, Dan Roth, Shih-Fu Chang, and Heng Ji. 2021. RESIN: A dockerized schema-guided cross-document cross-lingual cross-media information extraction and event tracking system. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations, NAACL-HLT*.

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 Additional Analysis of GENEVA

A.1 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.

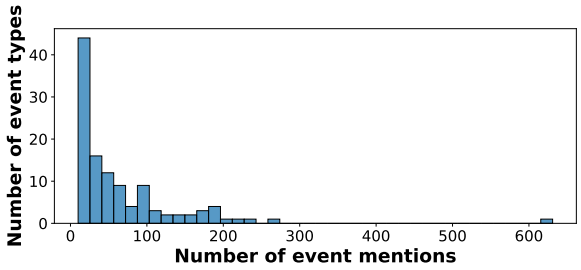

Figure 9: Distribution of event types by the number of event mentions in GENEVA.

A.2 Data Statistics for different benchmarking test suites

We present the data statistics for the various test suites in Table 7. For the training set of the low resource and few-shot test suites (indicated by $*$ in Table 7), we sample a smaller training set (as discussed in Section 4.3). 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 7) from this larger pool of data.

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

Table 7: 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.

A.3 Event Ontology Organization

The broad set of event types in GENEVA can be organized into a hierarchical structure of abstract event types. Adhering to the hierarchical tree structure introduced in MAVEN, we show the corresponding organization for event types in GENEVA in Figure 15. 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 15 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 15, we show the extent of the overlap of the mapped ACE events in the GENEVA event schema (text labels colored in red).[13] We can observe that although there is some 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.

A.4 Dataset Examples

We provide some examples of annotated sentences from the GENEVA dataset in Figure 10. We indicate the abstract event type in braces and cover an example from each abstraction.

B Automated Refinements for DEGREE

B.1 DEGREE

DEGREE 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. We illustrate DEGREE along with an example of its input prompt design in Figure 11.

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

Sentence	Event Type (Abstract)	Event Trigger	Arguments
Timothy McVeigh , who perpetrated the April 1995 bombing of the federal building in Oklahoma City , was not a member of a Patriot organization but identified strongly with the anti - government community of belief .	Committing_crime (Action)	perpetrated	Perpetrator: Timothy McVeigh Crime: the April 1995 bombing of the federal building in Oklahoma City
Canadian companies sent $28.5 billion in goods to the United States in February , up 1.6% from January revised level , while they imported $20.9 billion worth , up 2.4% .	Sending (Possession)	sent	Sender: Canadian companies Theme: $28.5 billion in goods Recipient: United States
With rail service in place and forty blocks of private property , it was ready to become a real town .	Becoming (Change)	become	Entity: it Final_category: a real town
Our police work ends , but our legal work begins .	Process_start (Scenario)	begins	Event: our legal work
The US administration calls for a total embargo on nuclear technology to Iran , and urges other nuclear suppliers , including the PRC , to take similar action (6847) .	Convincing (Sentiment)	urges	Speaker: The US administration Addressee: other nuclear suppliers , including the PRC Content: to take similar action

Figure 10: Illustration of example annotations from the GENEVA dataset for various different abstract types.

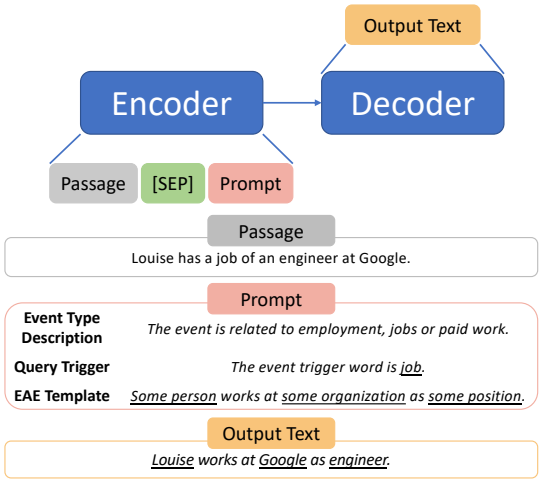

Figure 11: Model architecture of DEGREE (top half) and an illustration of a manually created prompt for the event type *Employment* (bottom half).

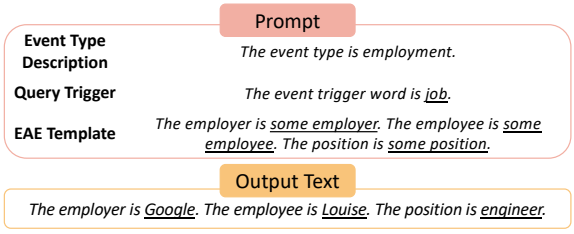

Figure 12: An illustration of an automatically generated prompt by DEGREE for the event type *Employment*.

Despite the superior performance of DEGREE in the low-data setting, it can not be directly 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 the wide set of events in GENEVA. Thus, there is a need to automate the manual human effort to scale up DEGREE.

B.2 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 DEGREE for these components below.

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 12.

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 11. For automating this mapping process, we propose a simple refinement of self-mapping, which 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 12.

Template Generation: The second subtask requires generating a natural sentence(s) using the argument role-mapped placeholder phrases (as shown in Figure 11). To automate this subtask, we create an event-agnostic template composed of argument role-specific sentences. For each argument role in the event, we generate a sentence of

	Original DEGREE	Automated DEGREE
ACE Dataset	73.5	72.7

Table 8: Model Performance in terms of F1 score for DEGREE and DEGREE on the ACE dataset.

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 12.

B.3 Ablation Study

In our work, we introduce automated refinements for scaling DEGREE for GENEVA. We provide an ablation study for these automated refinements (Automated DEGREE) on the ACE dataset in Table 8. We observe that the automated DEGREE almost at-par with DEGREE with a minor difference of only 0.8% F1 points.

C Impact of Pre-training

In this section, we explore the impact of pre-training models on the generalizability evaluation. We consider DEGREE and BERT_QA, pre-train them on the ACE dataset and show the model performance on low resource test suite in Figure 13.

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 as the number of training event mentions increases. In terms of zero-shot performance of the pre-trained models, DEGREE achieves a micro F1 score of 12.83% and BERT_QA achieves a score of 6.82% respectively. Poor zero-shot performance and diminishing performance gains indicate that GENEVA is distributionally distinct from ACE, which makes it challenging to achieve good model performance on GENEVA merely via transfer learning.

D Case Study: Is ACE diverse enough?

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) DEGREE pre-trained on the ACE dataset and (2) DEGREE 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 9.

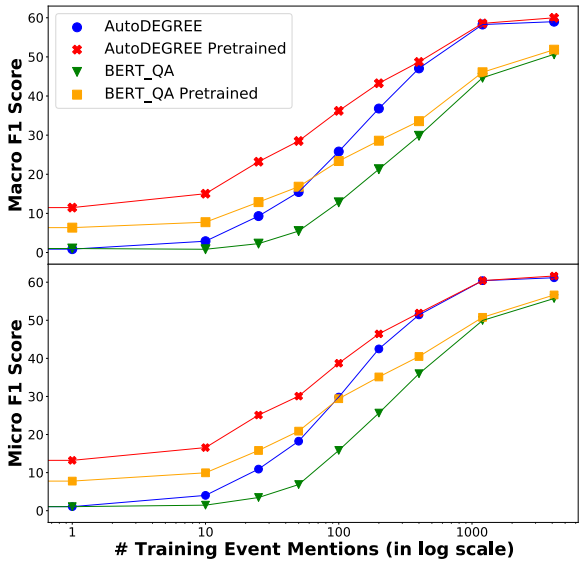

Figure 13: Model performance in macro F1 (top) and micro F1 (bottom) scores against the number of training event mentions (log-scale) for the low resource suite. Here we majorly compare the impact of pre-training on the model performance.

Abstract Event Type	Scratch Model	Pre-Trained Model	Δ
Action	35.48	38.93	3.45
Possession	45.65	50.63	4.98
Change	38.5	43.4	4.9
Sentiment	49.37	51.55	2.18
Scenario	30.87	34.59	3.72

Table 9: Model Performance in micro F1 on zero-shot with 10 event types split by abstract event types for (1) DEGREE with no pre-training (Scratch Model), and (2) Pre-Trained DEGREE on ACE (Pre-Trained Model). Δ: model performance difference.

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 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 well to out-of-domain event types. We also highlight the significance of GENEVA as its diverse evaluation setup helps analyze these trends.

ANNOTATION INSTRUCTIONS

Event: An event includes a class attribute with values such as occurrence, state, or reporting.
Event Arguments: Event participants that provide some event-centric information. We will try to make a sentence with frame element describing the event to consider it as a participant. If two frame elements are similar, we'll merge them and mark the more specific one as the event argument. We will remove frame elements which are super generic / present in most frames (e.g. Time, Place, Manner, Degree, Purpose, Explanation, Means); but we might want to include them if they are salient to the current event.
Entities: Some object in the world. We mark a frame element as entity if it's highly probable to be entity (and/or short noun phrase). We mark only those frame elements as entities which have been marked as event arguments.

Schema for frame element (FE) annotation (Is Event Argument?):
0: Not argument role 1: Argument role 2: Merge. Add name of parent FE. If multiple FEs with same name, mark only one as 1 and others as 2. Do not mark entity for these elements. 3: Ambiguous

Schema for entity annotations (Is Entity?): Only for frame elements mapped as argument roles
0: Not entity 1: Entity

Extra Tips:
Refer to examples if in doubt
Event Ontology Calibration Instructions:
- Flag if certain frames don't look like events
- Flag if certain frame merging looks incorrect based on argument roles

INPUT DATA				ANNOTATIONS		
Event Name	**Frame Name**	**Frame Element**	**Description**	**Is Event Argument?**	**Merged Argument**	**Is Entity?**
Institutionalizatio	Institutionalization		A Patient is committed to the care of a medical Facility by a proper Authority. Regardless of whether the Patient agrees or does not agree with their placement in the Facility, the Authority judges that it is in the best interest of the Patient or his environment that the Patient receive treatment in the Facility.			
		Time	The time interval during which the institutionalization happens.			
		Depictive	This FE describes a participant of the institutionalization as being in some state during the action.			
		Place	The location in which the Facility is situated.			
		Patient	The person who is committed to a facility with a view towards helping them mentally or physically.			

Figure 14: Figure illustrating the annotation process for EAE ontology creation. At the top, we present the annotation instructions. In the second bottom half, we show how the input data is presented along with fields for annotations.

E Human expert annotation for EAE ontology creation

Figure 14 present the annotation instructions and example input data for the human expert annotation process used for event argument ontology creation.

F Human validation for GENEVA

We provide an example of the annotation setup used for the *Ontology Quality Assessment* as part of GENEVA validation process in Table 10. Similarly, we provide the annotation setup and some examples for the *Annotation Comprehensiveness Assessment* in Table 11.

G Implementation Details

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

G.1 DEGREE

We closely follow the training setup by DEGREE for training the DEGREE models. We run experiments for DEGREE on a NVIDIA GeForce RTX 2080 Ti machine with support for 8 GPUs. We present the complete range of hyperparameter details in Table 12. We deploy early stopping criteria for stopping the model training.

G.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 DEGREE 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 13.

G.3 TANL

We report the hyperparameter settings for the TANL experiments in Table 14. 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.

G.4 DyGIE++

We report the hyperparameter settings for the DyGIE++ experiments in Table 15. Experiments for DyGIE++ were run on a NVIDIA GeForce RTX 2080 Ti machine with support for 4 GPUs.

G.5 OneIE

We report the hyperparameter settings for the OneIE experiments in Table 16. Experiments for OneIE were run on a NVIDIA GeForce RTX 2080 Ti machine with support for 4 GPUs.

G.6 Query&Extract

We report the hyperparameter settings for the Query&Extract experiments in Table 17. Experiments for OneIE were run on an NVIDIA GeForce RTX 2080 Ti machine with support for 4 GPUs.

G.7 TE

We use the original SRL engine and model provided in the repo for running the TE model. Since there was no training, we do not change any hyperparameters.

	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 10: Example for the human validation setup for ontology quality assessment.

Sentence	Event	Event Trigger	Annotated Arguments	Unannotated Arguments
The attackers were environmental terrorists upset about a new industry coming to town .	Attack	attackers	Assailant: environmental terrorists	Means, Victim, Weapon
United States Helps Uzbekistan Secure Dangerous Nuclear Materials : Energy agency announces completion of secret uranium transfer back to Russia	Assistance	Helps	Helper: United States Goal: Secure Dangerous Nuclear Materials	Benefited_party, Focal_entity, Means

Table 11: Examples for the human validation setup for annotation comprehensiveness assessment.

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	200
Max Output Length	150

Table 12: Hyperparameter details for DEGREE model.

H 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 Figures 16 and 17 respectively. Figure 18 displays the results for the zero-shot and cross-type transfer settings.

PLM	BERT-Large
Training Batch Size	12
Eval Batch Size	8
Learning Rate	1×10^{-5}
# Training Epochs	50
# Evaluations per Epoch	5
Max Sequence Length	300
Max Answer Length	50
N-Best Size	20

Table 13: Hyperparameter details for BERT_QA model.

PLM	T5-Base
Training Batch Size	8
Eval Batch Size	12
Learning Rate	5×10^{-4}
# Training Epochs	20
Evaluation per # Steps	100
Max Sequence Length	256
# Beams	8

Table 14: Hyperparameter details for TANL model.

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 15: Hyperparameter details for DyGIE++ model.

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 16: Hyperparameter details for OneIE model.

PLM	BERT-Large
Training Batch Size	16
Eval Batch Size	16
Learning Rate	5×10^{-5}
Weight Decay	0.001
# Training Epochs	5
Evaluation per # Epoch	10
Entity Embedding Size	100

Table 17: Hyperparameter details for Query&Extract model.

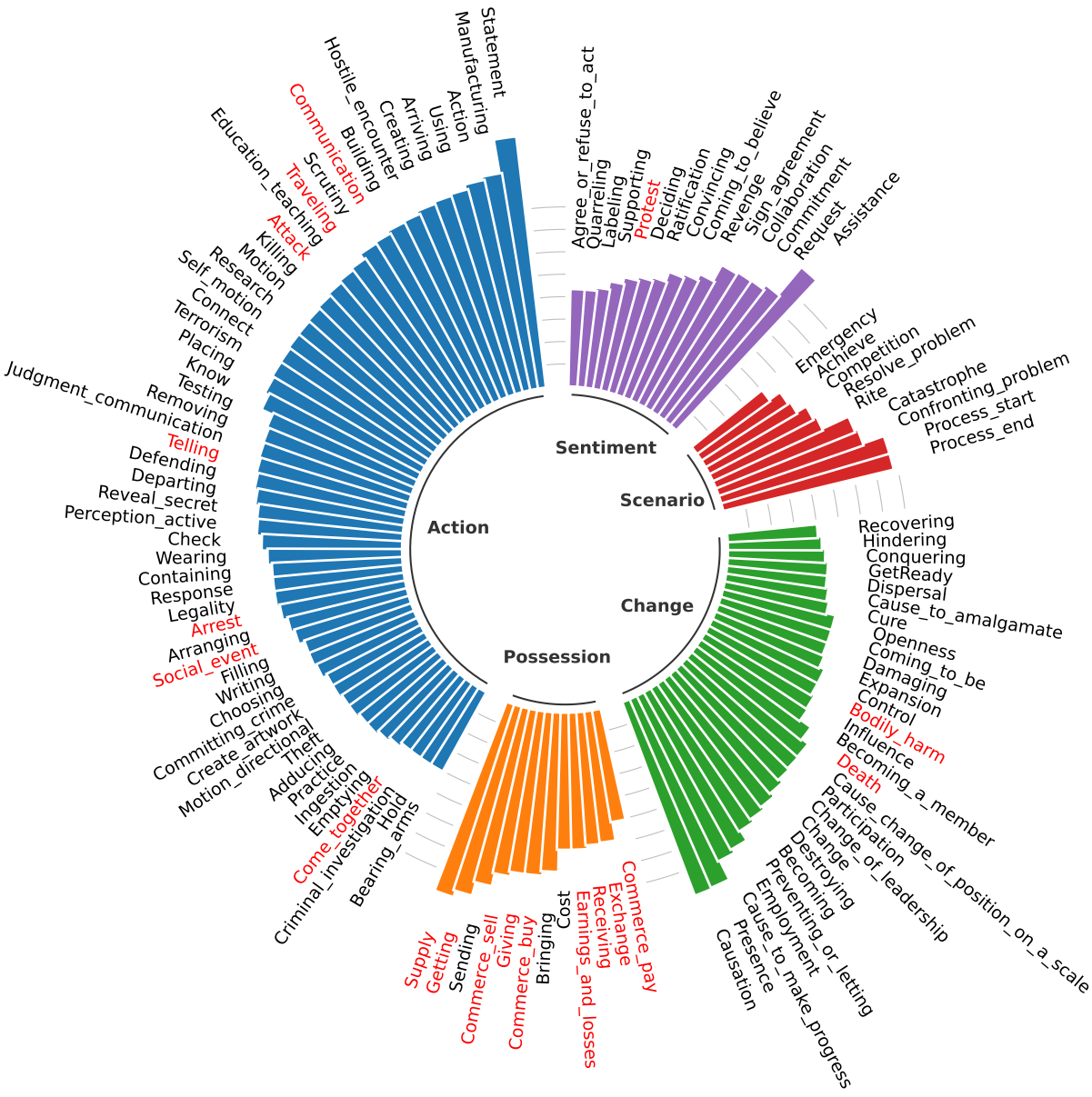

Figure 15: Circular bar plot for the various event types present in the GENEVAdataset organized into abstract event types. The height of each bar is proportional to the number of event mentions for that event (height is in log-scale). Bar labels colored in red are the set of overlapping event types mapped from the ACE dataset.

Model	LR-10		LR-25		LR-50		LR-100		LR-200		LR-400		LR-1200		Full Training	
	Micro	**Macro**	**Micro**	**Macro**	**Micro**	**Macro**	**Micro**	**Macro**	**Micro**	**Macro**	**Micro**	**Macro**	**Micro**	**Macro**	**Micro**	**Macro**
DEGREE	2.64	2.53	12.43	9.98	23.47	23.95	28.31	26.79	41.57	39.14	51.80	50.81	61.35	60.85	62.19	62.26
	2.02	1.76	10.69	8.41	22.50	18.71	32.06	28.65	43.34	42.47	51.82	49.89	58.98	59.82	61.23	61.46
	2.13	1.78	13.48	12.21	21.04	19.59	32.53	30.55	44.27	43.89	44.85	41.81	61.49	61.12	62.44	62.12
	5.77	6.67	7.96	6.68	14.99	13.59	31.89	30.14	40.94	37.58	51.42	50.15	62.38	62.29	61.90	61.14
	2.33	1.67	13.86	12.45	23.09	19.09	37.03	34.91	43.79	43.03	49.84	49.41	60.69	59.72	62.51	62.98
	2.98	**2.88**	**11.68**	**9.95**	**21.02**	**18.99**	**32.36**	**30.21**	**42.78**	**41.22**	**49.95**	**48.41**	**60.98**	**60.76**	**62.05**	**61.99**
BERT_QA	3.16	1.32	5.23	4.02	12.67	11.95	19.92	17.39	22.77	19.63	29.87	26.80	49.64	46.42	53.91	52.58
	1.89	1.24	6.89	5.41	11.79	8.70	17.99	14.32	26.96	23.76	34.50	30.49	48.51	46.46	54.14	52.57
	0.80	0.26	3.33	1.89	11.13	9.25	16.65	13.06	26.79	24.28	36.25	32.38	50.95	49.67	53.88	52.30
	4.53	3.74	5.92	5.06	7.70	6.27	16.94	13.19	21.82	19.08	30.83	26.08	49.03	46.18	54.20	52.60
	1.13	0.96	8.76	5.95	9.38	7.34	15.30	13.29	27.24	23.84	33.70	30.31	48.61	46.01	55.40	53.29
	2.30	**1.50**	**6.03**	**4.47**	**10.53**	**8.70**	**17.36**	**14.25**	**25.12**	**22.12**	**33.03**	**29.21**	**49.35**	**46.95**	**54.31**	**52.67**
TANL	1.74	0.91	6.59	2.54	8.76	4.51	22.19	13.25	28.93	17.60	48.66	36.43	64.52	52.74	70.97	64.32
	0.82	0.71	3.21	1.77	17.84	9.31	20.69	12.19	29.78	18.83	50.10	36.93	65.27	55.38	72.69	67.17
	0.80	0.12	4.88	2.86	6.67	4.20	13.50	7.31	40.37	28.72	47.98	33.00	66.27	57.65	71.49	65.51
	1.84	0.90	4.12	2.59	6.96	4.06	16.04	8.36	32.63	18.32	50.11	36.19	64.14	54.03	71.78	65.09
	1.43	0.28	6.30	3.06	14.87	7.78	22.91	12.07	23.16	13.53	50.23	36.53	66.64	55.83	70.79	64.55
	1.33	**0.58**	**5.02**	**2.56**	**11.02**	**5.97**	**19.07**	**10.64**	**30.97**	**19.40**	**49.42**	**35.82**	**65.37**	**55.13**	**71.54**	**65.33**
DyGIE++	0.00	0.00	0.91	0.60	1.80	2.21	5.65	4.68	13.39	8.77	28.73	18.95	57.74	46.13	65.88	56.28
	0.11	0.04	0.39	0.13	4.22	2.56	5.90	4.42	13.98	9.87	31.69	22.18	56.18	47.37	66.53	56.52
	0.01	0.01	0.84	0.35	1.61	1.22	5.98	5.08	17.53	12.24	27.32	18.08	58.07	47.40	66.26	55.28
	0.06	0.22	0.19	0.14	1.93	1.91	9.70	5.44	13.56	8.02	29.79	18.20	56.73	47.26	65.29	54.95
	0.00	0.00	0.00	0.00	4.28	1.79	7.64	5.21	14.49	9.31	28.21	18.72	57.11	43.63	65.31	54.21
	0.04	**0.05**	**0.47**	**0.24**	**2.77**	**1.94**	**6.97**	**4.97**	**14.59**	**9.64**	**29.15**	**19.23**	**57.17**	**46.36**	**65.85**	**55.45**
OneIE															30.02	22.44
															30.53	22.14
															29.16	21.00
															30.12	22.82
															30.30	22.74
															30.03	**22.23**
Query and Extract															40.70	32.25
															40.52	32.04
															40.02	32.47
															40.41	**32.25**

Figure 16: 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. LR-XX denotes low resource with XX training mentions.

Model	FS-1		FS-2		FS-3		FS-4		FS-5	
	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro
DEGREE	28.69	29.11	44.58	45.08	47.46	49.95	52.69	55.38	53.71	57.41
	29.73	31.38	41.07	43.34	47.48	52.40	50.96	54.73	54.09	55.88
	31.73	31.94	40.61	44.10	49.39	51.03	51.13	54.67	53.82	57.53
	30.65	30.84	43.62	46.46	49.34	51.88	51.51	55.07	52.90	58.26
	29.77	32.23	44.99	45.74	48.72	51.26	50.39	53.91	55.61	60.05
	30.11	**31.10**	**42.97**	**44.94**	**48.48**	**51.30**	**51.34**	**54.75**	**54.03**	**57.83**
BERT_QA	13.87	14.11	23.53	24.84	26.89	28.63	28.30	29.88	37.09	40.72
	13.77	14.01	24.31	25.72	21.25	25.03	31.72	34.55	41.29	41.03
	16.10	16.26	20.97	22.15	23.35	24.53	32.51	34.34	36.31	39.35
	18.20	16.93	25.30	25.61	30.33	31.14	34.09	36.24	34.17	37.85
	14.99	15.43	24.62	26.19	29.87	30.82	32.08	33.88	37.65	41.37
	15.39	**15.35**	**23.75**	**24.90**	**26.34**	**28.03**	**31.74**	**33.78**	**37.30**	**40.06**
TANL	17.50	12.61	32.73	27.16	44.67	45.55	47.12	45.41	49.36	51.94
	15.10	13.11	37.06	34.01	46.09	46.70	40.41	37.66	48.21	46.86
	19.91	16.78	34.67	33.42	43.55	41.00	53.49	53.66	54.90	56.89
	17.93	15.78	30.28	27.48	35.01	31.56	51.19	50.87	56.16	56.66
	14.46	11.31	31.00	27.79	44.19	44.20	51.69	52.05	51.37	54.63
	16.98	**13.92**	**33.15**	**29.97**	**42.70**	**41.80**	**48.78**	**47.93**	**52.00**	**53.40**
DyGIE++	5.01	7.86	10.77	15.87	17.79	23.91	23.96	31.37	25.80	36.13
	4.87	8.25	12.13	19.03	18.42	27.58	23.01	32.84	24.80	35.95
	4.66	7.82	10.30	15.51	16.33	22.76	22.22	31.73	25.73	32.38
	6.14	9.03	14.24	18.11	17.62	24.68	20.00	26.77	27.04	36.39
	5.20	8.40	12.90	18.53	17.06	25.85	22.28	29.48	27.31	38.33
	5.18	**8.27**	**12.07**	**17.41**	**17.44**	**24.96**	**22.29**	**30.44**	**26.14**	**35.84**

Figure 17: 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. FS-X denotes few-shot with X training mentions per event.

Model	ZS-1		ZS-5		ZS-10		CTT	
	Micro	Macro	Micro	Macro	Micro	Macro	Micro	Macro
DEGREE	24.66	24.97	34.42	33.36	41.11	40.49	28.78	30.59
	24.91	24.94	35.23	34.79	38.74	40.21	27.51	30.57
	23.60	22.90	36.50	35.18	40.42	40.07	28.26	30.53
	24.69	23.91	33.25	32.88	38.03	37.89	27.16	30.28
	22.42	23.18	34.02	33.72	38.86	39.30	27.78	30.16
	24.06	**23.98**	**34.68**	**33.99**	**39.43**	**39.59**	**27.90**	**30.43**
BERT_QA	4.86	2.92	17.96	15.05	22.80	19.27	12.82	11.94
	4.76	2.79	24.35	19.94	25.87	22.97	11.56	10.79
	5.72	3.95	19.17	14.95	28.22	25.61	12.59	11.06
	3.88	2.50	23.16	19.23	23.13	18.88	14.60	14.19
	6.02	3.84	23.02	18.53	21.20	19.26	4.30	3.98
	5.05	**3.20**	**21.53**	**17.54**	**24.24**	**21.20**	**11.17**	**10.39**

Figure 18: Complete set of results of the 5 different runs for all models for the zero-shot (ZS) and cross-type transfer (CTT) test suite. Here Micro is the micro F1 score and Macro is the macro F1 score. ZS-X denotes zero-shot with X training events.