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

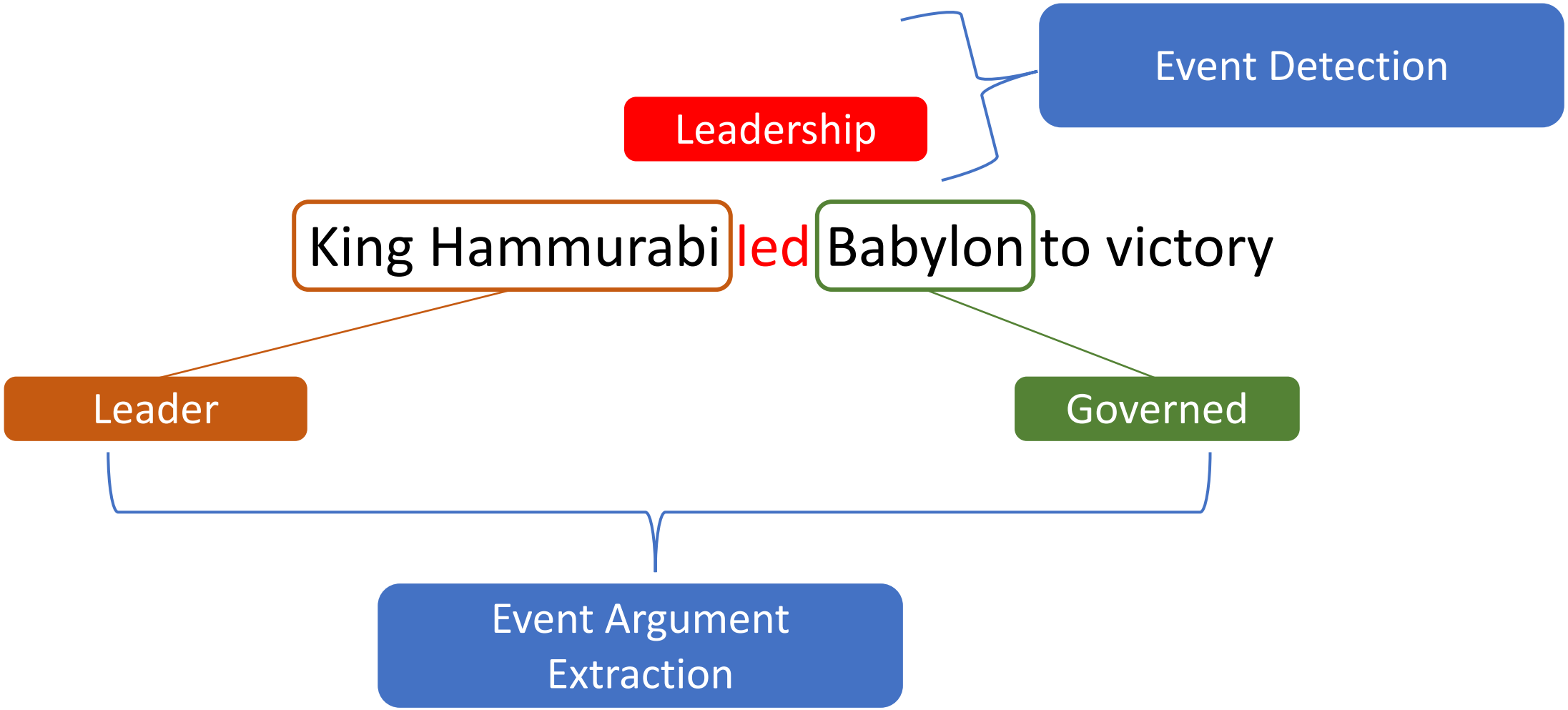
Outline

- Introduction
- Dataset
- Methodology
- Experiments and Results
- Conclusion and Future Work

What is Event Extraction?

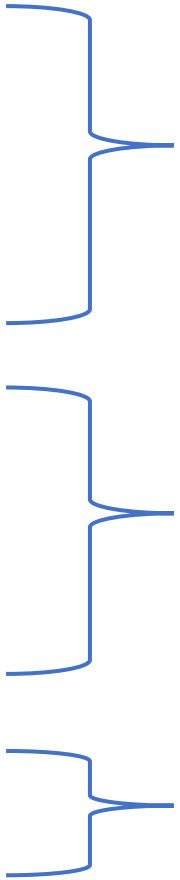
King Hammurabi led Babylon to victory

What is Event Extraction?



Existing EAE Benchmarking Datasets

- ACE-05
- ERE
- RAMS
- WikiEvents
- MAVEN



Limited number of events and arguments

Limited diversity in event types

Limited to ED

New diverse EAE dataset GENEVA + four benchmarking test suites for testing generalizability

Introducing AutoDEGREE – generalizable and robust EAE model

Thorough generalizability evaluation of various existing EAE models

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

Outline

- Introduction
- Dataset
- Methodology
- Experiments and Results
- Conclusion and Future Work

FrameNet

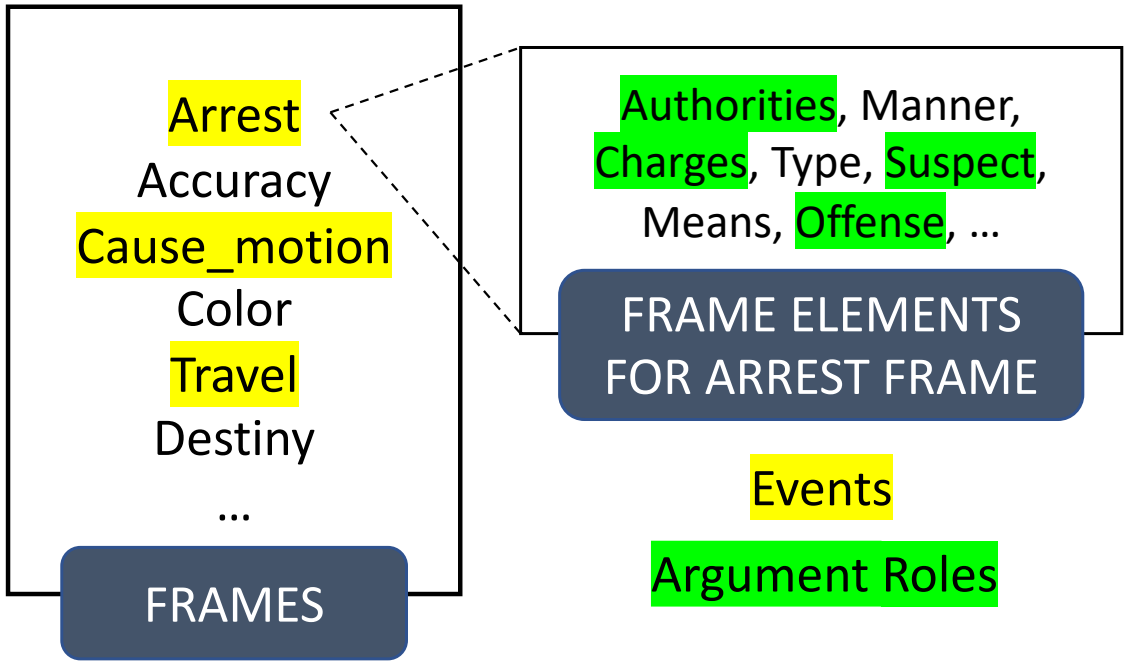
- Large set of linguistically human-annotated data of frames following the Frame Semantic Theory

- Over 1200+ Semantic Frames (Potential Events)
  - Lexical Units (Event Triggers)
  - Frame elements (Event Arguments)
  - Frame relations (Argument relations)

Framenet is too fine-grained. All Frames are not Events

Complex frame structure. Not all elements are arguments

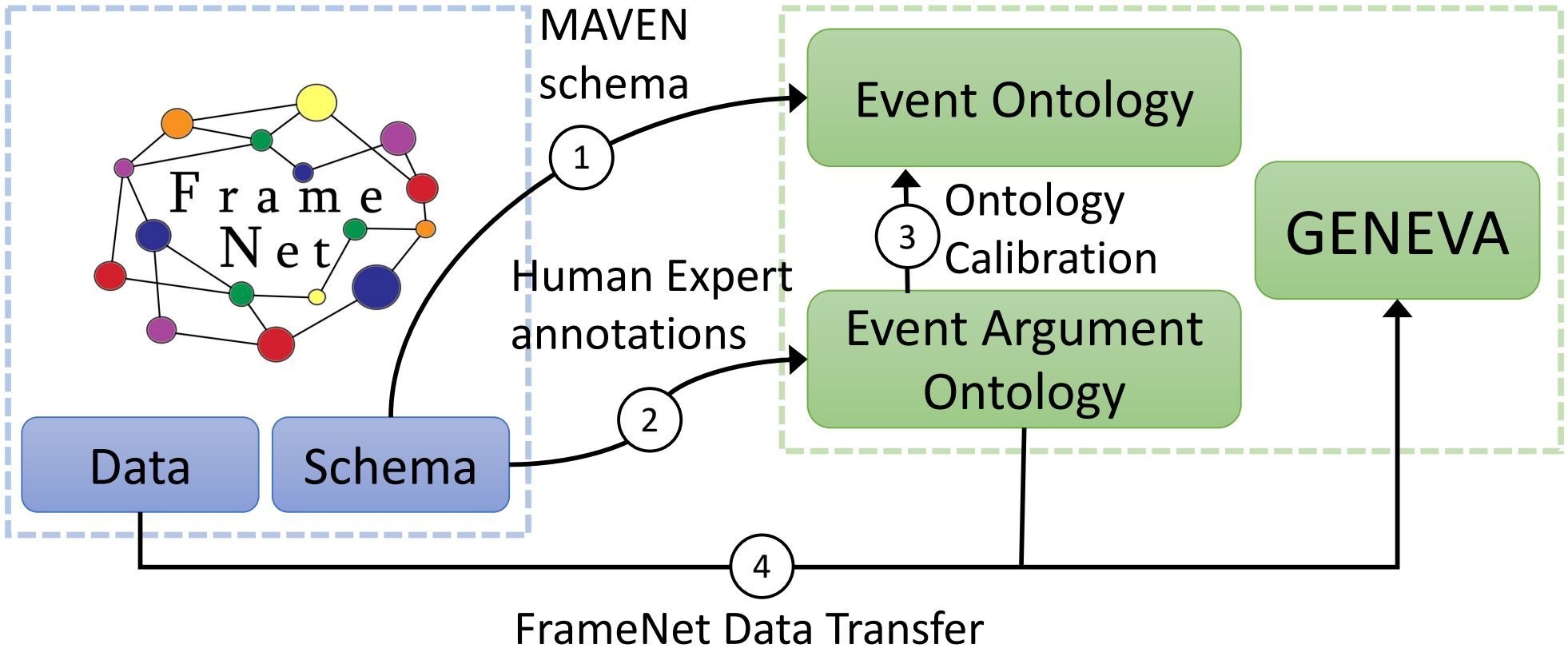
FrameNet



Framenet is too fine-grained.
All Frames are not Events

Complex frame structure. Not
all elements are arguments

GENEVA


MAVEN
schema
1
Event Ontology
Ontology
Calibration
3
GENEVA
Human Expert
annotations
2
Event Argument
Ontology
Data
Schema
4
FrameNet Data Transfer

Our created ontology

	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%

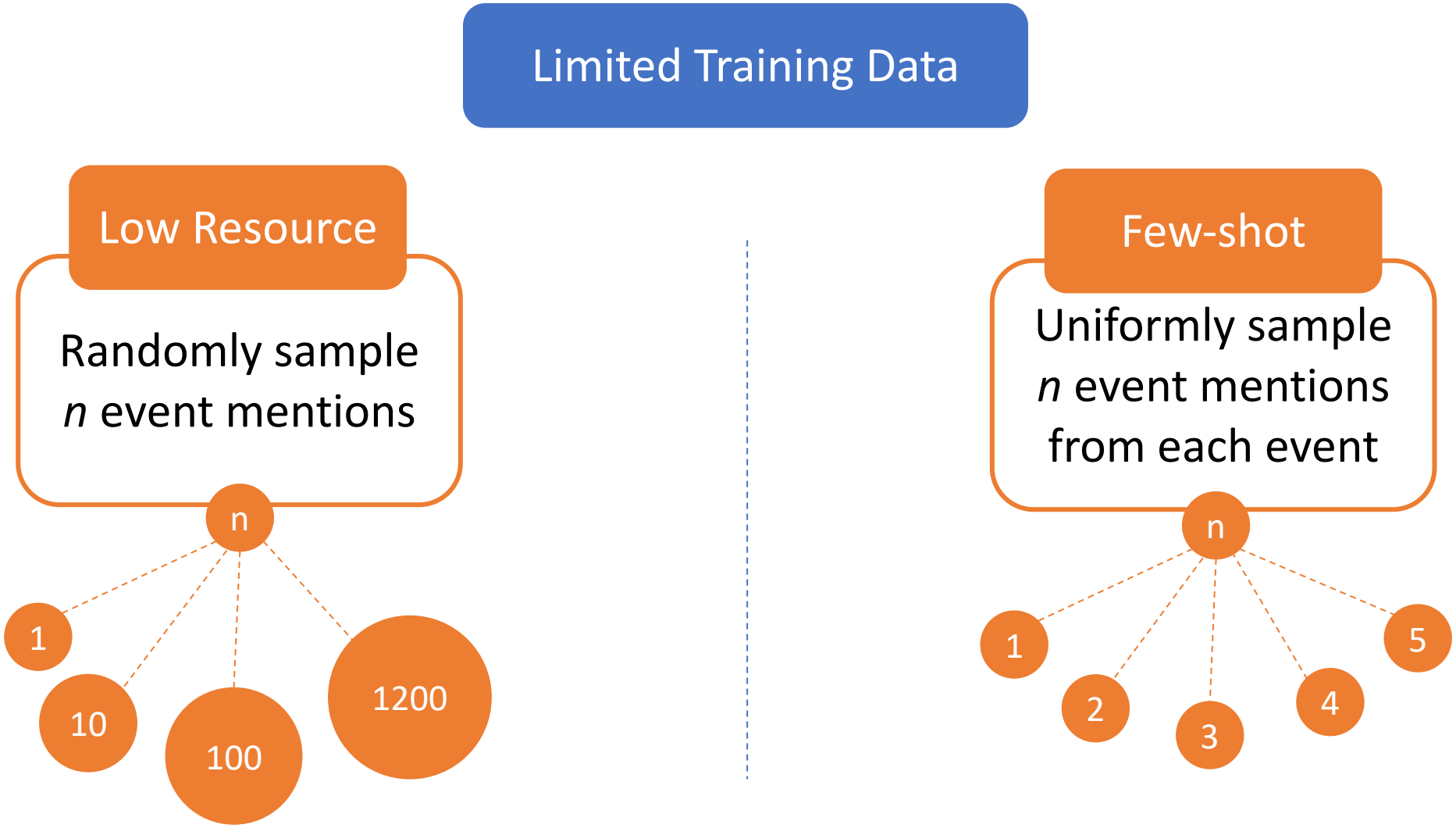
Table 1: Full and GENEVA ontology Statistics. AR = Argument Role. An ontology covers an abstract type if it has 5+ events of that abstract type. Entity AR refers to argument roles that are entities.

Large coverage of event types and argument roles

Diverse set of abstract event types covered

Cover non-entity arguments which aren't covered before

Benchmarking Setup

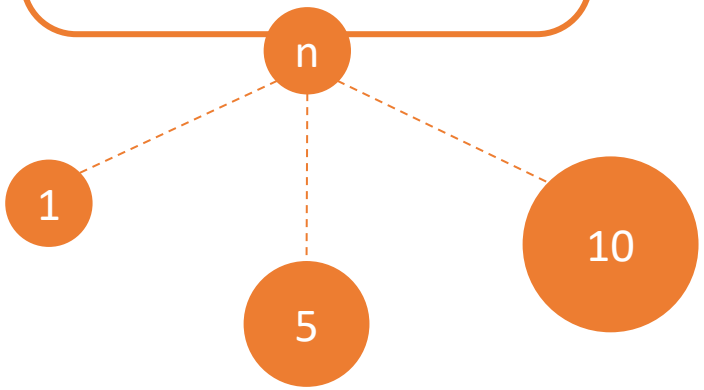


Benchmarking Setup

Unseen Event Data

Zero-shot

Training data only from *n* event types



Cross-Type Transfer

Training data from all events of an abstract type

Data Statistics

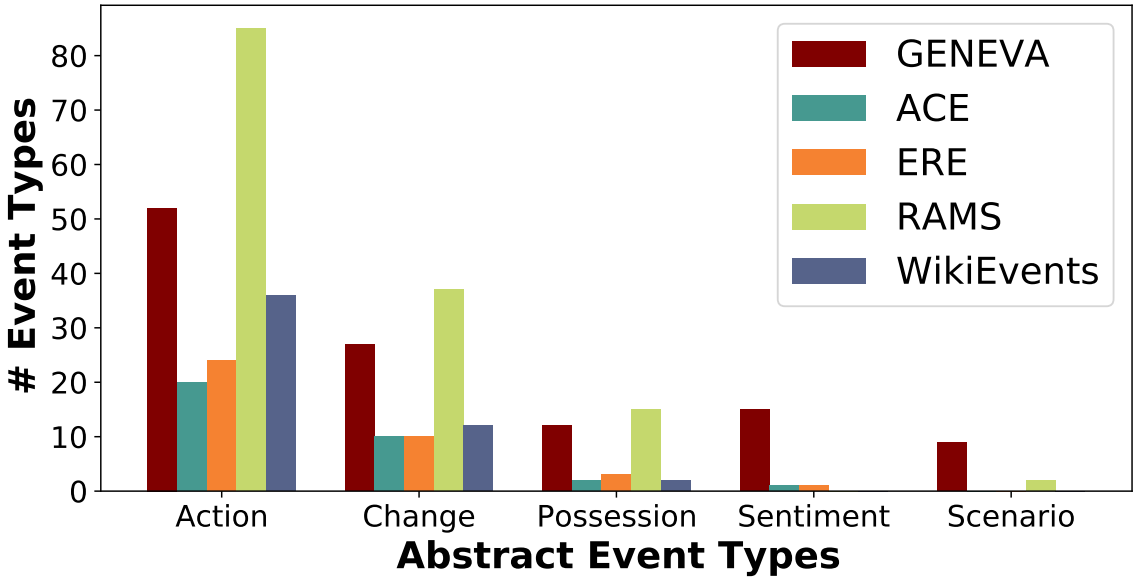
High coverage of event types and argument roles

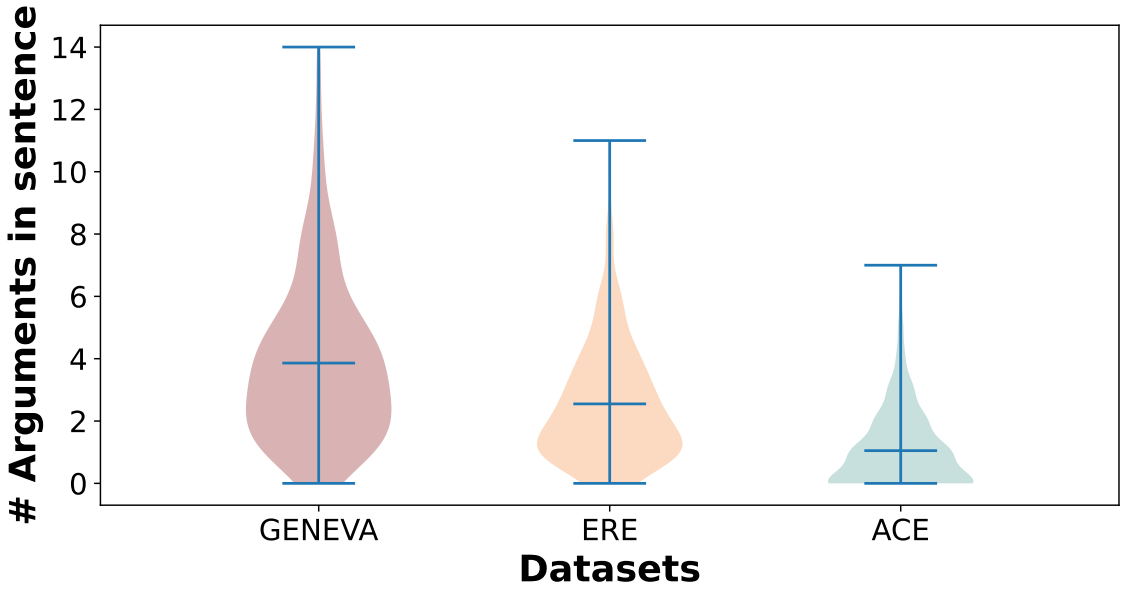
Challenging: Average mentions per event type and argument role is less

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.

Data Statistics



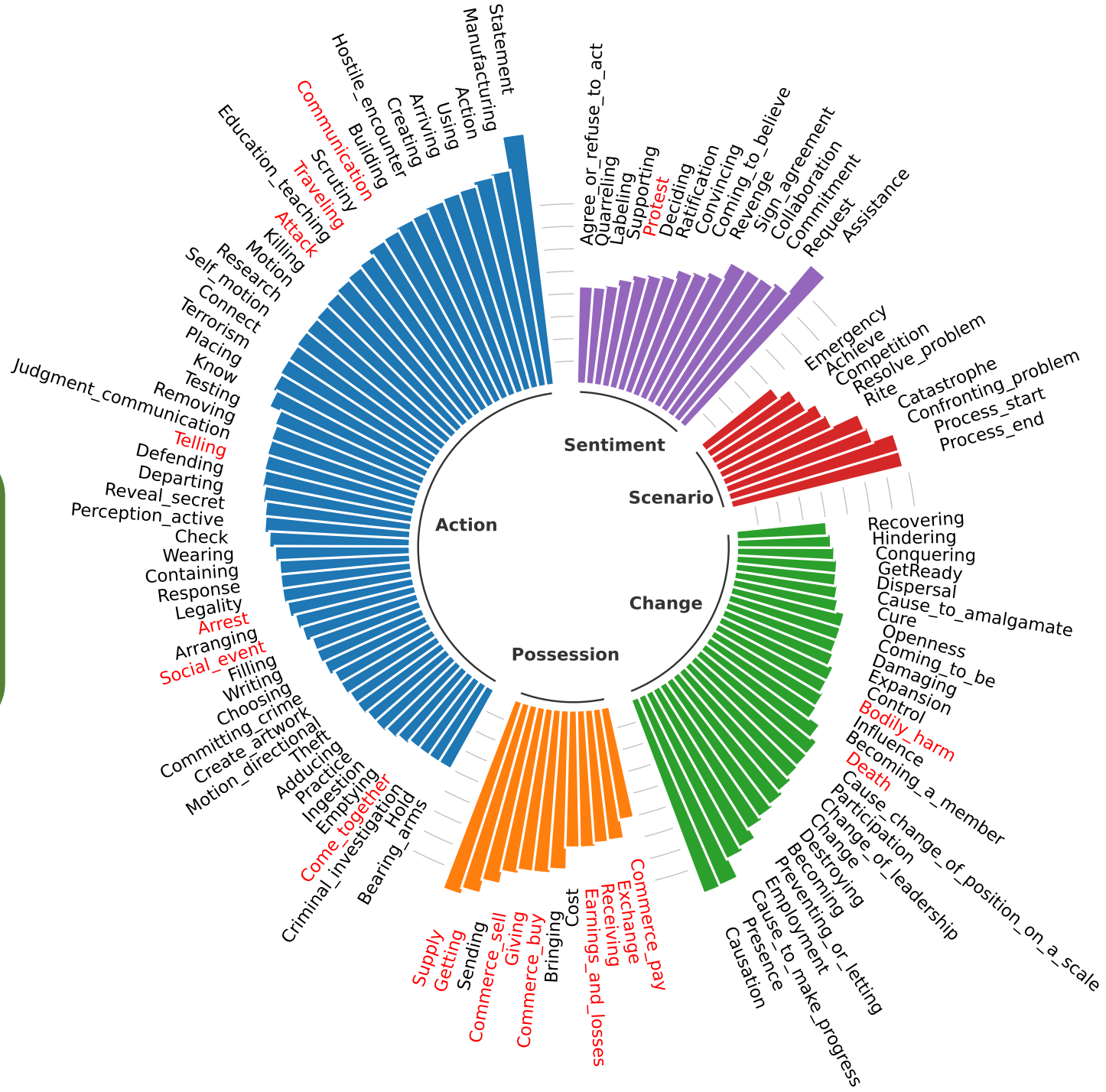


GENEVA is more diverse than all other existing datasets

GENEVA has more argument role mentions per sentence

Data Statistics

GENEVA coverage based on abstract types
ACE (red) only covers a part of the event ontology



Outline

- Introduction
- Dataset
- Methodology
- Experiments and Results
- Conclusion and Future Work

EAE Models

- OneIE
- DyGie++
- QAEE
- DEGREE
- …

Classification-based models

Do not work for new event types
Can't do zero-shot or few-shot!

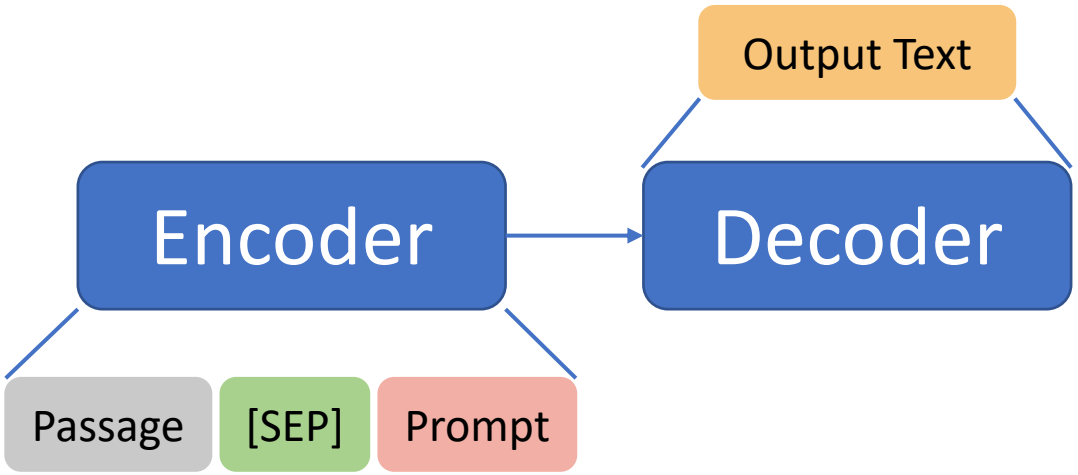
Generative approaches

Require human annotation for each new event. Can't scale up!

Towards scaling up DEGREE

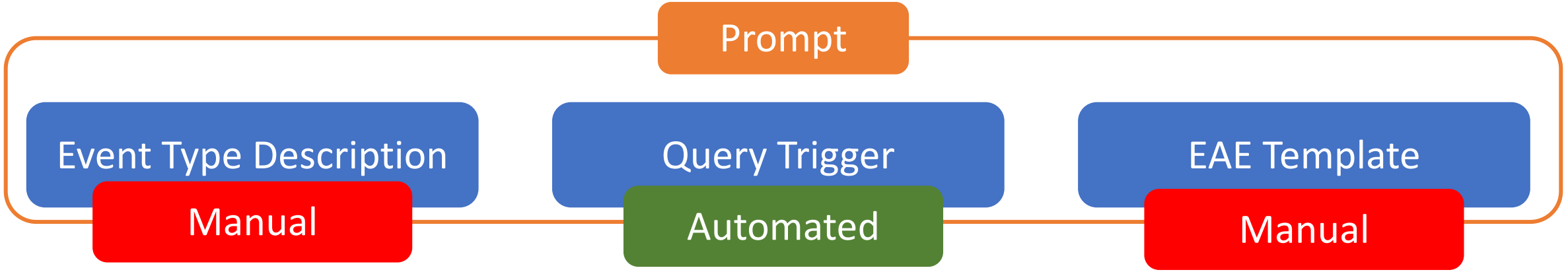
DEGREE Model - EAE

Passage: *Earlier Monday , a 19-year-old Palestinian riding a bicycle detonated a 30-kilo (66-pound) bomb near a military jeep in the Gaza Strip , injuring three soldiers.*



Prompt for DEGREE(EAE)	
◆ Event Type Description	*The event is related to conflict and some violent physical act.*
★ Query Trigger	*The event trigger word is detonated.*
☾ EAE Template	*some attacker attacked some facility, someone, or some organization by some way in somewhere.*
Output Text	
Palestinian attacked jeep and soldiers by bomb in Gaza Strip.	

DEGREE Model - EAE



DEGREE – Automating Event Description

Event Type Description

The event is related to conflict or some violent physical act.

Event Type Description

The event type is conflict.

DEGREE – Automating Template

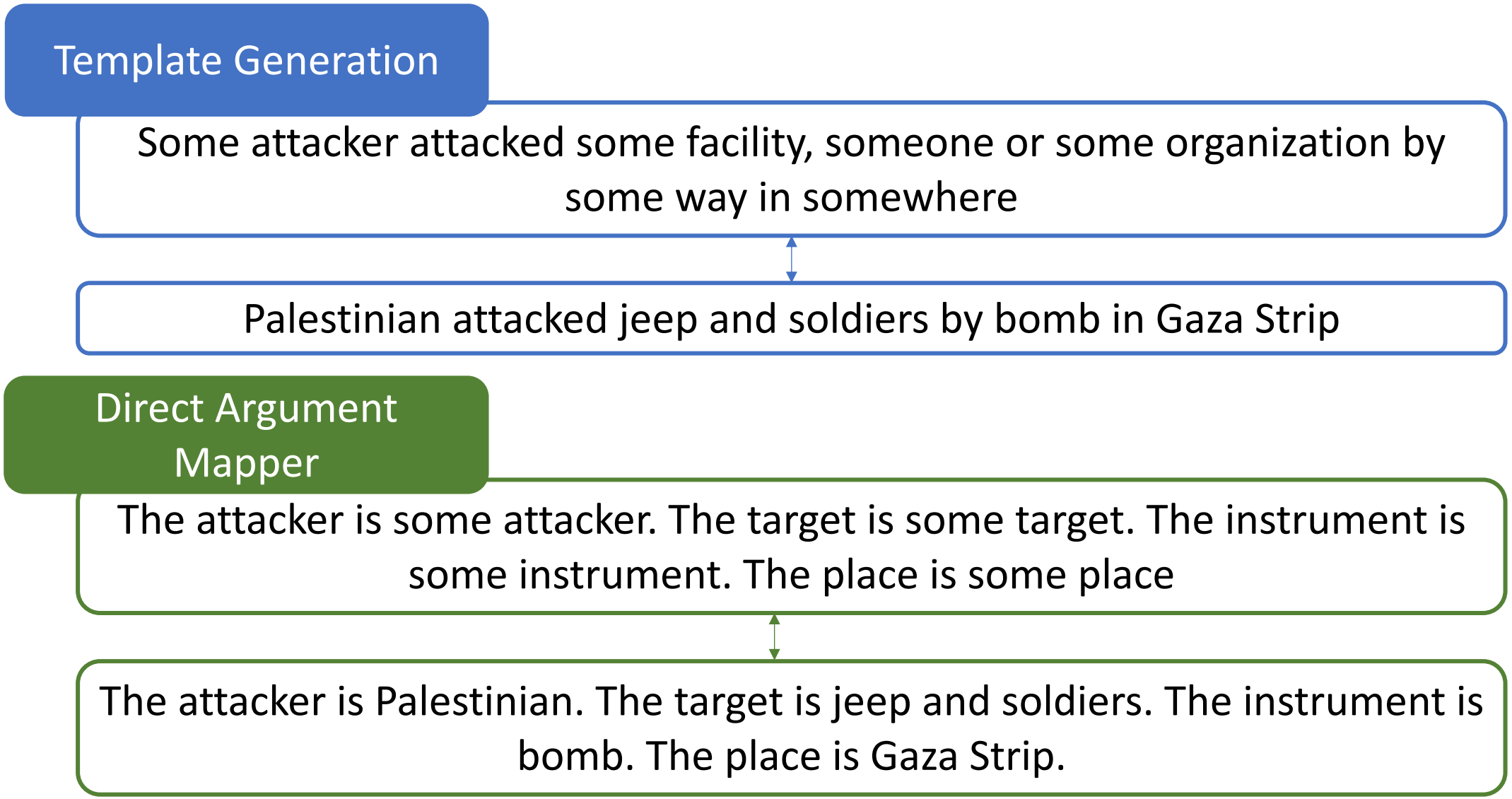
Role Mapping

Attacker -> Some attacker | Target -> Some facility, someone or some org …

Argument Mapping

Attacker -> some attacker | Target -> some target …

DEGREE – Automating Template



Outline

- Introduction
- Dataset
- Methodology
- Experiments and Results
- Conclusion and Future Work

Experimental Setup

Baselines

OneIE

DyGIE++

QAEE

TANL

Query & Extract

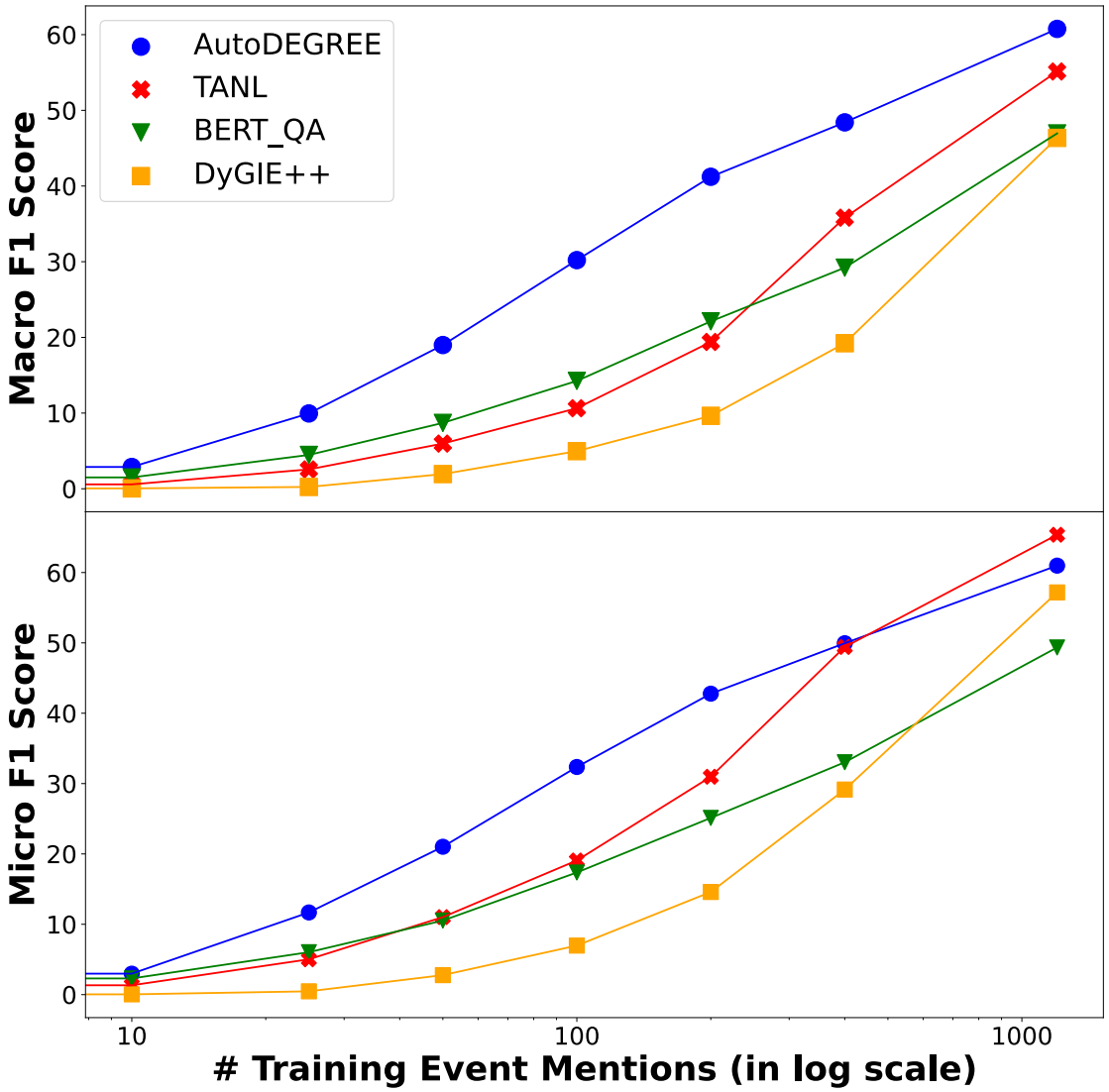
Evaluation

Micro-F1

Macro-F1

We want to generalize; hence a good F1 score across wide range of events

Results – Low Resource Benchmark

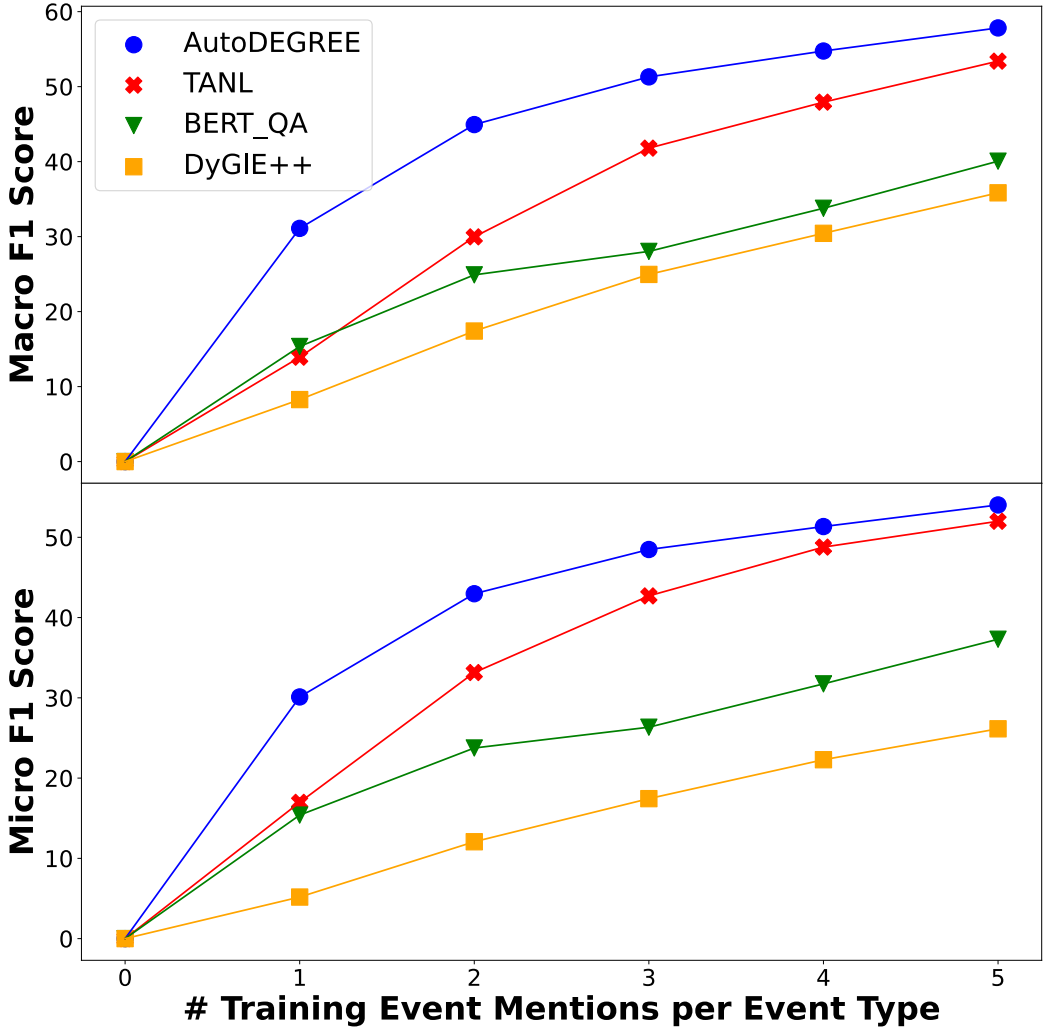


OneIE and Query & Extract achieve poor overall scores and not included here

DEGREE performs best across both metrics and in all data settings

TANL and DyGIE++ give good micro F1 for higher events but poor macro F1 indicating poor generalizability

Results – Few-shot Benchmark



AutoDEGREE model outperforms all other baseline models

Traditional classification methods show poor performance

Results – Unseen Data Setting

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

TANL and DyGIE++ show poor zero-shot performance and not included here

AutoDEGREE model outperforms all other baseline models

Analysis – GENEVA v/s ACE

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

Both DEGREE and BERT_QA perform better on ACE and relatively poorly on GENEVA benchmarks

Analysis – GENEVA v/s ACE

	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

Breakdown shows that non-entity arguments are more difficult to extract and shows the additional challenge introduced by GENEVA dataset

Outline

- Introduction
- Dataset
- Methodology
- Experiments and Results
- Conclusion and Future Work

Conclusion

- Using similarity of SRL and EAE, we constructed a new vast EAE ontology spanning 115 event types and 220 argument roles
- Utilizing this ontology, we construct a new generalizability benchmarking dataset GENEVA comprising four test suites.
- We benchmark various existing EAE models on our benchmarking test suites and inspire further research on generative models for EAE.
- Analysis further shows how GENEVA poses new challenges for EAE models and we anticipate future generalizability benchmarking efforts.