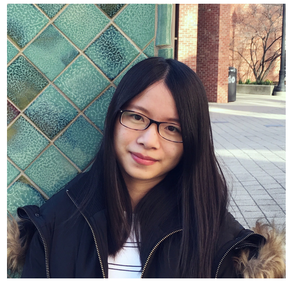
Instruction-Tuning LLMs for Event Extraction with Annotation Guidelines

Saurabh Srivastava*, Sweta Pati*, Ziyu Yao
George Mason University





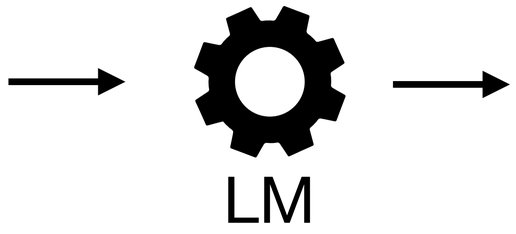
Code Format Facilities Event Extraction

- Wang, Li, and Ji (2023) found that LMs perform better when they read/write the structured event information in code format

```
# This is an event extraction task ...
# The following lines describe the task definition
@dataclass
class Extradite(JusticeEvent):
    mention
    agent
    person
    destination

# This is the text to analyze
text = After getting caught they were transferred to
       the U.S. for trial.
```

Event Schema




```
Extradite(
  mention = “transferred”,
  person = [“they”],
  destination = [“U.S.”]
)
```

(Img source: Srivastava et al., 2025)

LMs were pre-trained to read and write code;

Structured knowledge representation: easier to represent hierarchical schema, type-argument association, type constraints, etc.

Wang, Xingyao, Sha Li, and Heng Ji. "Code4Struct: Code Generation for Few-Shot Event Structure Prediction." *ACL* 2023.

Annotation Guidelines Serve as Helpful Hints

- Saniz et al. (2024) further discovered that the expert-written annotation guidelines, formatted as Python docstrings and comments, are helpful hints
  - Mainly verified in Named Entity Recognition

```
# The following lines describe the task definition
@dataclass
class ProgrammingLanguage(Entity):
    """Refers to a programming language used in the development of AI
    applications and research. Annotate the name of the programming
    language, such as Java and Python."""

    span: str  # Such as: "Java", "R", "CLIPS", "Python", "C + +"

@dataclass
class Metric(Entity):
    """Refers to evaluation metrics used to assess the performance of AI
    models and algorithms. Annotate specific metrics like F1-score."""

    span: str  # Such as: "mean squared error", "DCG", ...
```

Our work: Does including guidelines help the more challenging Event Extraction task?

Sainz, Oscar, et al. "GoLLIE: Annotation Guidelines improve Zero-Shot Information-Extraction." *ICLR 2024*.

Event Extraction w/ Guidelines in Code Format

```
# This is an event extraction task ...
# The following lines describe the task definition
@dataclass
class Extradite(JusticeEvent):
"""The event is triggered by the formal request and subsequent transfer of an individual from one state or country to another for legal reasor
indicative of this event type, not 'Transport' which involves general movement without legal context."""
    mention  # The text span that triggers the event.
    agent  # The agent plays a crucial role in the extradition process, often being a legal or governmental body.
    person # Examples are 'she', 'him', 'her'. The person is the individual being extradited.
    destination  # Examples are 'jurisdiction', 'Hague', 'state'. The destination is the place to which the person is being extradited.
# This is the text to analyze
text = After getting caught they were transferred to the U.S. for trial.
```

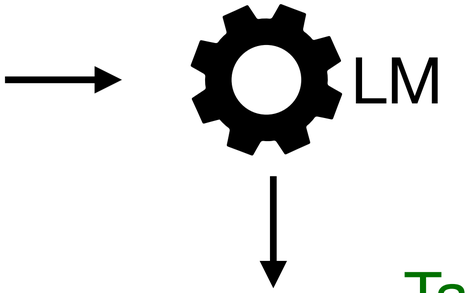
- Problem setup: instruction tuning (FLAN/Google Research 2022)

Wei, Jason, et al. "Finetuned Language Models are Zero-Shot Learners." *ICLR 2022*.

Event Extraction w/ Guidelines in Code Format

Instruction

```
# This is an event extraction task ...
# The following lines describe the task definition
@dataclass
class Extradite(JusticeEvent):
"""The event is triggered by the formal request and subsequent transfer of an in
indicative of this event type, not 'Transport' which involves general movement w
    mention    # The text span that triggers the event.
    agent  # The agent plays a crucial role in the extradition process, often l
    person # Examples are 'she', 'him', 'her'. The person is the individual bei
    destination  # Examples are 'jurisdiction', 'Hague', 'state'. The destina
# This is the text to analyze
text = After getting caught they were transferred to the U.S. for trial.
```



Target to maximize

```
Extradite(
  mention = “transferred”,
  person = [“they”],
  destination = [“U.S.”]
)
```

- Problem setup: instruction tuning (FLAN/Google Research 2022)
- Additionally,
  - Can machines generate more effective guidelines?
  - How do guidelines help in low-data settings (i.e., small amounts of training examples)?

Wei, Jason, et al. "Finetuned Language Models are Zero-Shot Learners." *ICLR 2022*.

Automatic Generation of Annotation Guidelines

- Expert-written guidelines are not always available and may not be the most helpful to LMs
- Can SOTA LMs (e.g., GPT-4o) *reverse-engineer* guidelines from data?

Guideline Generation Prompt (Guideline-PN)

You are an expert in annotating NLP datasets for event extraction. Your task is to generate annotation guidelines for the event type Extradite which is a child event type of super class JusticeEvent.

The event schema is as follows:

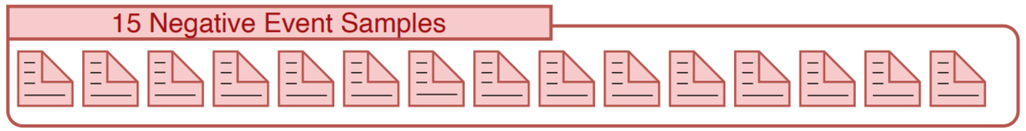
```
@dataclass
class Extradite(JusticeEvent):
    mention
    (...)
```

The below examples are positive examples, as they match the Event Type being annotated

10 Positive Event Samples

The following examples are negative examples, as they illustrate different event types provided for contrast and differentiation:

15 Negative Event Samples



Instructions
1. Identify and List All Unique Arguments.
2. Define the Event Type: Write 5 clear and specific definitions, starting with "The event is triggered by ...":
3. Define Each Argument:** For each argument, provide 5 definitions.

More Task Instructions (...)

Variants:

- Guideline-Positive (P)
- Guideline-Positive+Negative (PN)
- Guideline-Positive+Siblings (PS)

(More variants in paper)

Automatic Generation of Annotation Guidelines

Examples of Annotation Guidelines for Event Type: Extradite (ACE05)	
GUIDELINE-H Avg. Length - 107.67 tokens	**Event Type:** An EXTRADITE Event occurs whenever a PERSON is sent by a state actor from one PLACE to another place for the purposes of legal proceedings there. **Arguments:** - AGENT: The extraditing agent. - PERSON: The person being extradited.
GUIDELINE-P Avg. Length - 163.87 tokens	**Event Type:** The Extradition event refers to the formal process where one jurisdiction delivers a person accused **(...)** The event can be triggered by terms such as 'extradition' **(...)** Edge cases include situations where the term 'extradition' is used metaphorically or in a non-legal context. **Arguments:** - AGENT:**(...)** the agent is the organization or authority **(...)**. Examples include 'court', 'government', **(...)** - PERSON: **(...)** individual who is being transferred to another jurisdiction. Examples are 'she', **(...)**
GUIDELINE-PN Avg. Length - 285.24 tokens	**Event Type:** The event is triggered by the formal request **(...)** for legal reasons. Triggers such as 'extradition' are indicative of this event type, not 'Transport' which involves general movement without legal context. **Arguments:** - AGENT: The agent is responsible for the legal and procedural aspects of the extradition,**(...)**. An example is 'the original court' **(...)** - PERSON: **(...)** one who is being moved from one place to another under legal authority. For example, 'he' **(...)**

(Notation: distinctions from other event types, example mentions, and edge cases)

Does including guidelines help EE?

- Datasets: ACE05 (Doddington et al., 2004) and RichERE (Song et al., 2015)
- Model: Llama-3.1-8B (Meta GenAI, 2024)
- Two settings: w/o or w/ Negative Sampling (NS)
  - NS: 15 negative samples per training example (15x more training examples)

Full training set: 16k and 9k

Experiments	ACE w/o NS				ACE w/ NS				RichERE w/o NS				RichERE w/ NS			
	TI	TC	AI	AC	TI	TC	AI	AC	TI	TC	AI	AC	TI	TC	AI	AC
NoGuideline	39.57	39.57	31.05	29.73					35.11	35.11	27.16	25.32				
Guideline-H	40.71	40.71	30.76	28.64					–	–	–	–				
Guideline-P	**51.46**	**51.46**	**37.82**	**35.20**					34.38	34.38	28.04	26.35				
Guideline-PN	49.60	49.60	35.80	32.81					**40.89**	**40.89**	**30.04**	**27.18**				
Guideline-PS	47.93	47.93	37.19	34.88					32.41	32.41	24.63	22.78				

Guidelines help; Machine Guidelines > Human Guidelines;

Does including guidelines help EE?

- Datasets: ACE05 (Doddington et al., 2004) and RichERE (Song et al., 2015)
- Model: Llama-3.1-8B (Meta GenAI, 2024)
- Two settings: w/o or w/ Negative Sampling (NS)
  - NS: 15 negative samples per training example (15x more training examples)

Full training set: 16k and 9k

Experiments	ACE w/o NS				ACE w/ NS				RichERE w/o NS				RichERE w/ NS			
	TI	TC	AI	AC	TI	TC	AI	AC	TI	TC	AI	AC	TI	TC	AI	AC
NoGuideline	39.57	39.57	31.05	29.73	**84.15**	**84.15**	**64.99**	**61.96**	35.11	35.11	27.16	25.32	42.27	42.27	32.38	31.56
Guideline-H	40.71	40.71	30.76	28.64	56.30	56.30	44.82	43.13	–	–	–	–	–	–	–	–
Guideline-P	**51.46**	**51.46**	**37.82**	**35.20**	72.86	72.86	55.01	53.73	34.38	34.38	28.04	26.35	67.92	67.92	52.29	44.93
Guideline-PN	49.60	49.60	35.80	32.81	80.77	80.77	63.20	60.34	**40.89**	**40.89**	**30.04**	**27.18**	75.35	75.35	**60.85**	**57.10**
Guideline-PS	47.93	47.93	37.19	34.88	79.23	79.23	59.00	56.88	32.41	32.41	24.63	22.78	**76.45**	**76.45**	60.42	56.26

Guidelines help; Machine Guidelines > Human Guidelines;

Guidelines may or may not complement Negative Sampling

How do guidelines help in low-data settings?

- Hypothesis: guidelines gain more in low-data settings with the extracted data heuristics

Training size: 2k

Experiments	ACE w/o NS				ACE w/ NS				RichERE w/o NS				RichERE w/ NS			
	TI	TC	AI	AC	TI	TC	AI	AC	TI	TC	AI	AC	TI	TC	AI	AC
NoGuideline	10.60	10.60	5.19	3.68	31.64	31.64	25.91	24.22	19.87	19.87	13.34	11.69	36.29	36.29	28.15	25.58
Guideline-H	29.01	29.01	16.37	14.78	32.62	32.62	25.35	22.87	–	–	–	–	–	–	–	–
Guideline-P	36.91	36.91	24.17	21.24	56.99	56.99	**43.44**	**40.51**	**40.28**	**40.28**	**21.97**	**18.33**	62.04	62.04	46.33	42.03
Guideline-PN	30.94	30.94	19.27	17.64	**60.29**	**60.29**	42.88	39.95	31.23	31.23	19.48	17.51	**67.16**	**67.16**	**47.85**	**43.39**
Guideline-PS	**40.53**	**40.53**	**28.03**	**26.12**	55.1	55.1	41.57	38.91	26.16	26.16	16.64	15.19	58.95	58.95	42.79	38.1

2k training w/ Guidelines + NS outperforms full-size training

Full training set: 16k

Experiments	ACE w/o NS			
	TI	TC	AI	AC
NoGuideline	39.57	39.57	31.05	29.73

Full training set: 9k

RichERE w/o NS				RichERE w/ NS			
TI	TC	AI	AC	TI	TC	AI	AC
35.11	35.11	27.16	25.32	42.27	42.27	32.38	31.56

How do guidelines help in low-data settings?

- Hypothesis: guidelines gain more in low-data settings with the extracted data heuristics

Training size: 100

	ACE w/ NS				RichERE w/ NS			
	TI	**TC**	**AI**	**AC**	**TI**	**TC**	**AI**	**AC**
NoGuide	**37.08**	**37.08**	**21.53**	**19.18**	24.98	24.98	15.05	13.15
H	29.00	29.00	17.93	16.34	–	–	–	–
P	27.95	27.95	15.94	14.21	23.93	23.93	13.56	12.71
PN	29.60	29.60	17.87	15.92	27.43	27.43	**17.10**	**15.28**
PS	29.85	29.85	19.49	17.04	19.61	19.61	11.77	10.48

But when the training size is too small, LMs cannot be effectively instruction-tuned to utilize the guidelines

More details in our paper!

- After an LM is tuned to utilize the task instruction, will it also gain more **generalization** ability?
  - Yes! See our cross-schema generalization results
- Do the guidelines help **smaller** LMs? **Non-Llama** LMs?
  - Yes! We observed similar patterns w/ Llama-3.2-1B and Qwen2.5-Coder-1.5B
- What event types benefit the most from guidelines?
  - **Frequent and less frequent event types both benefit from guidelines**, but for extremely infrequent event types, LMs cannot use guidelines well