61 ACL 2023

💬SPEECH: Structured Prediction with Energy-Based Event-Centric Hyperspheres

Preprint: http://arxiv.org/abs/2305.13617, Project: https://github.com/zjunlp/SPEECH

Shumin Deng[1], Shengyu Mao[2], Ningyu Zhang[2] ✉, Bryan Hooi[1] ✉

[1]National University of Singapore & NUS-NCS Joint Lab, Singapore

[2]Zhejiang University & AZFT Joint Lab for Knowledge Engine, China

1 

2 

Toronto, Canada
July 9-14, 2023

July, 2023

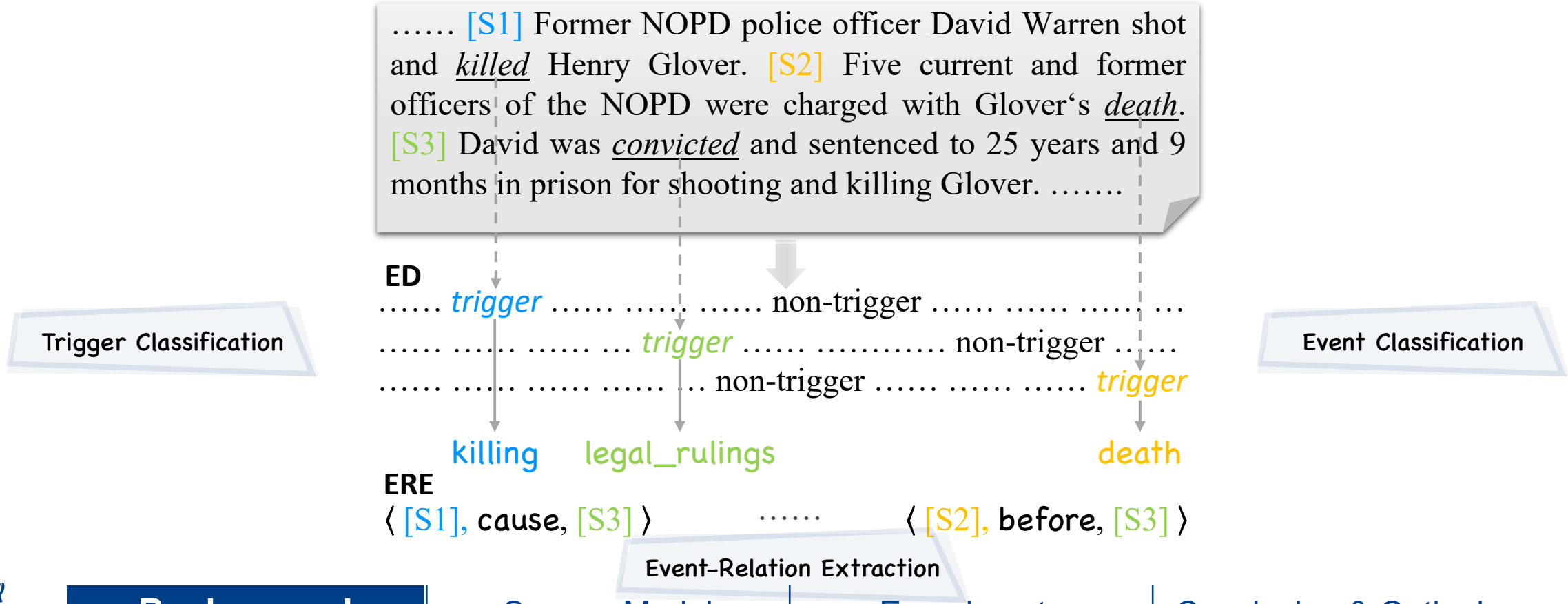
CONTENTS

CONTENTS

SEEKING TRUTH
PURSUING INNOVATION

Event-Centric Structured Prediction

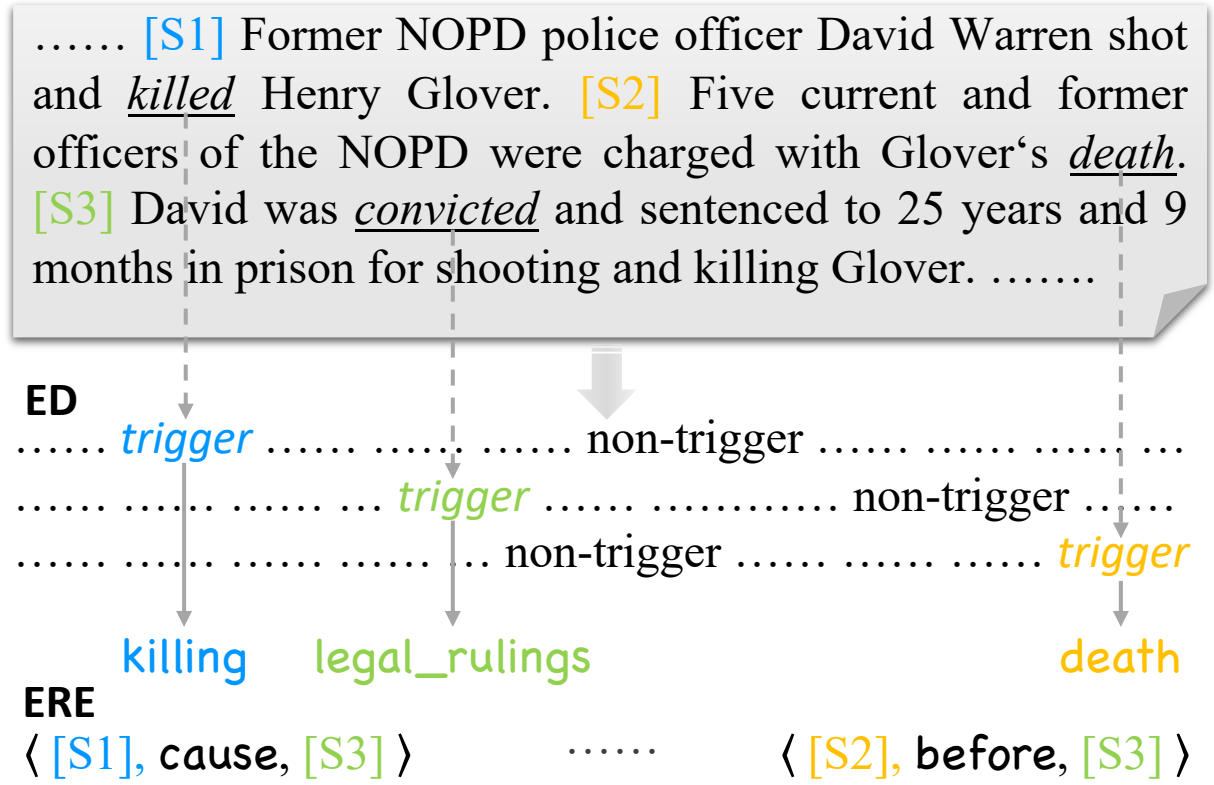
- **Structured Prediction**: predicted outputs are complex structured components
- **Event-Centric Structured Prediction**: considers manifold structures and dependency of events from texts, including intra-/inter-sentence structures, e.g., **ED** (**Event Detection**), **ERE** (**Event-Relation Extraction**)



Challenging Problems

How to model manifold event structures

How to efficiently represent these events



CONTENTS

Preliminaries (Energy Function)

- **Structured Prediction Energy Networks (SPENs)**
  - Input: $\boldsymbol{x} \in \mathcal{X}$; Model: $\mathbf{M}_\Phi$; Structured Outputs: $\mathbf{M}_\Phi(\boldsymbol{x}) \in \tilde{\mathcal{Y}}$
  - score structured outputs with an **energy function** $E_\Theta : \mathcal{X} \times \tilde{\mathcal{Y}} \to \mathbb{R}$
    - parameterized by Θ that iteratively optimize the energy between the input/output pair
- **Event-Centric Structured Prediction (ECSP) with Energy Networks**
  - Given a feature vector $\boldsymbol{x}$ belonging to one of T labels
  - The model output: $\mathbf{M}_\Phi(\boldsymbol{x}) = \{0,1\}^T \in \tilde{\mathcal{Y}}$
  - **energy function** $E_\Theta(\boldsymbol{x}, \boldsymbol{y}) = E_\Theta^{local}(\boldsymbol{x}, \boldsymbol{y}) + E_\Theta^{label}(\boldsymbol{y})$

$$= \sum_{i=1}^{T} y_i V_i^\top f(\boldsymbol{x}) + w^\top g(W\boldsymbol{y})$$

After learning the energy function, prediction minimizes energy

$$\tilde{\boldsymbol{y}} = \arg\min_{\boldsymbol{y} \in \tilde{\mathcal{Y}}} E_\Theta(\boldsymbol{x}, \boldsymbol{y})$$

  - final theoretical optimum for SPENs:

$$\min_\Theta \max_\Phi \sum \left[\triangle\left(\mathbf{M}_\Phi(\boldsymbol{x}_i), \boldsymbol{y}_i\right) - E_\Theta\left(\boldsymbol{x}_i, \mathbf{M}_\Phi(\boldsymbol{x}_i)\right) + E_\Theta\left(\boldsymbol{x}_i, \boldsymbol{y}_i\right) \right]_+$$

$[a]_+ = \max(0, a)$

$\triangle(\tilde{\boldsymbol{y}}, \boldsymbol{y})$: hinge loss

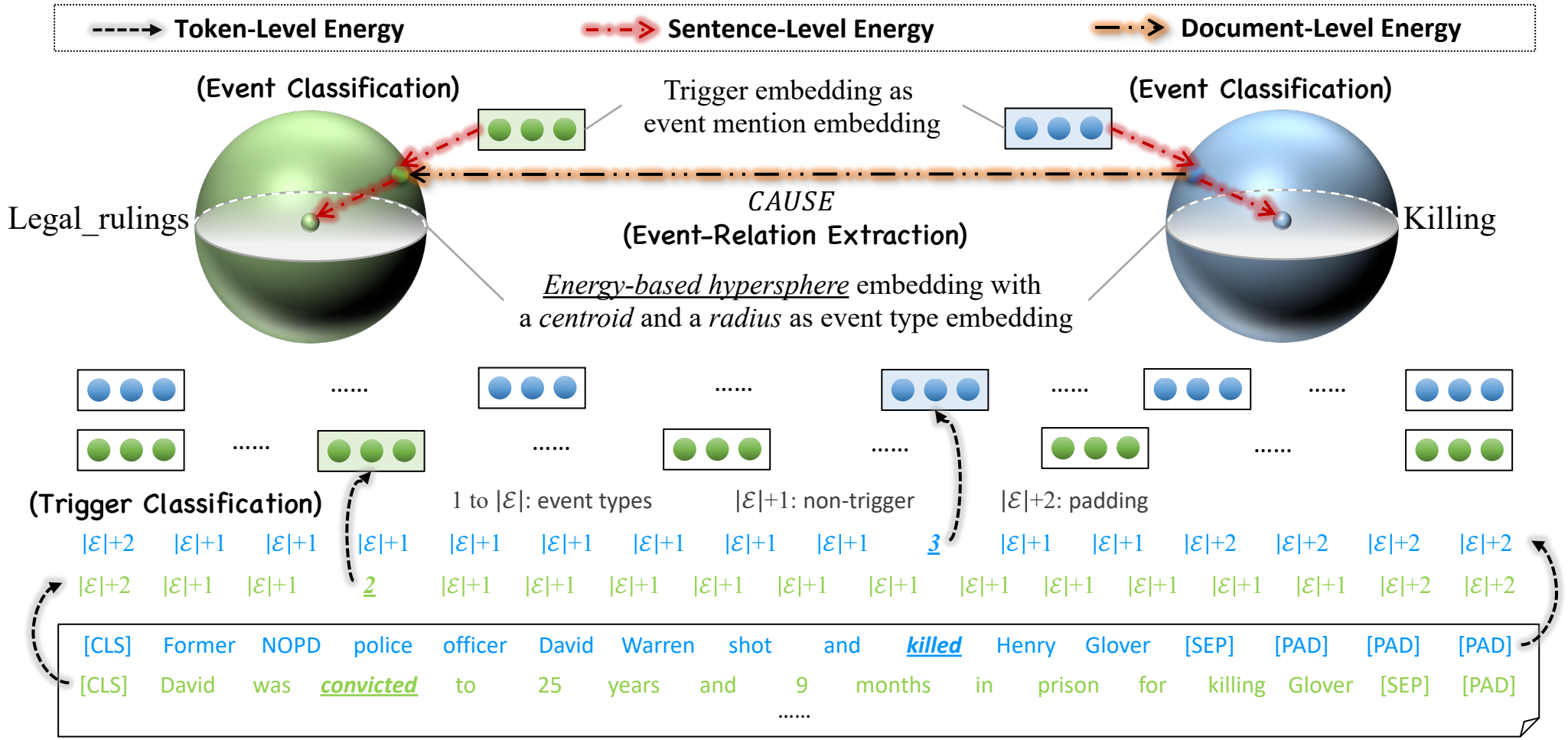
Structured Prediction Energy Networks (ICML 2016); Learning Approximate Inference Networks for Structured Prediction (ICLR 2018)

Problem Formulation

- Dataset: $\mathcal{D} = \{\mathcal{E}, \mathcal{R}, \mathcal{X}\}$
  - event class set: $\mathcal{E} = \{e_i \mid i \in [1, |\mathcal{E}|]\}$
  - event-relation set: $\mathcal{R} = \{r_i \mid i \in [1, |\mathcal{R}|]\}$
  - event mentions: $\mathcal{X} = \{\boldsymbol{X}_i \mid i \in [1, K]\}$ $\boldsymbol{x} = \{\boldsymbol{x}_j \mid j \in [1, L]\}$ a token sequence
- Tasks:
  - Trigger Classification: $\tilde{\boldsymbol{y}} = \mathbf{M}_\Phi(\boldsymbol{x})$
    - Find the index $t (1 \leq t \leq L)$ for the trigger $\boldsymbol{x}_t$
    - Categorize $\boldsymbol{x}_t$ into a specific event class $e_i \in \mathcal{E}$
  - Event Classification: $\tilde{\boldsymbol{Y}} = \mathbf{M}_\Phi(\boldsymbol{X})$
    - Predict the event label e_i for each event mention $\boldsymbol{X}_i$
  - Event-Relation Extraction: $\tilde{\boldsymbol{z}} = \mathbf{M}_\Phi(\ddot{\boldsymbol{X}})$
    - Identify the relation $r_i \in \mathcal{R}$ for a pair of event mentions $\ddot{\boldsymbol{X}}_{\langle ij \rangle} = (\boldsymbol{X}_i, \boldsymbol{X}_j)$

Model Overview

SPEECH: **S**tructured **P**rediction with **E**nergy-Based **E**vent-**C**entric **H**yperspheres



Three-Level Energy for Three Tasks

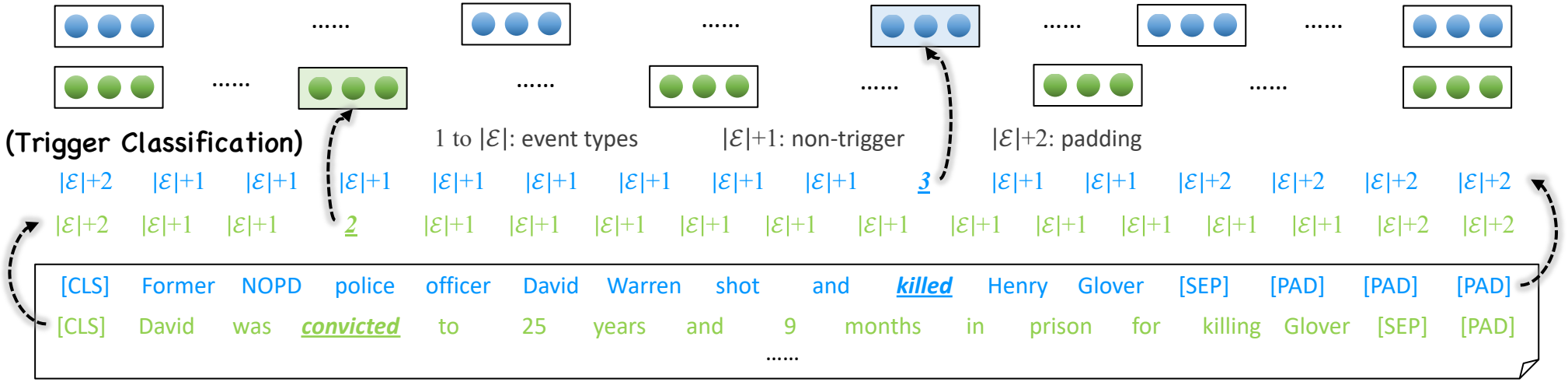
❑ Trigger Classification

❑ Token-level Energy Function

$$E_\Theta(\boldsymbol{x}, \boldsymbol{y}) = -\left(\sum_{n=1}^{L} \sum_{i=1}^{|\mathcal{E}|+2} \underbrace{y_n^i \left(V_{1,i}^\top f_1(\boldsymbol{x}_n) \right)}_{local} + \sum_{n=1}^{L} \underbrace{y_{n-1}^\top W_1 y_n}_{label} \right)$$

❑ Loss Function

$$\mathcal{L}_{tok} = \sum_{i=1}^{L} \left[\triangle(\tilde{\boldsymbol{y}}_i, \boldsymbol{y}_i) - E_\Theta(\boldsymbol{x}_i, \tilde{\boldsymbol{y}}_i) + E_\Theta(\boldsymbol{x}_i, \boldsymbol{y}_i) \right]_+ + \mu_1 \mathcal{L}_{\mathrm{CE}}(\tilde{\boldsymbol{y}}_i, \boldsymbol{y}_i)$$



Three-Level Energy for Three Tasks

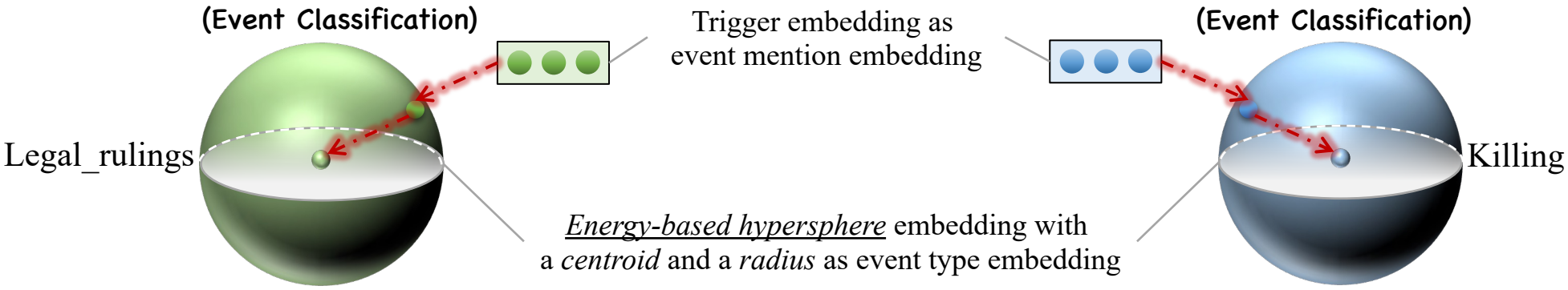
Event Classification

- Hyperspherical Measurement Function $\mathcal{S}(\boldsymbol{X}, \mathcal{P}_i) = \frac{\exp^{-[\,\|\mathcal{P}_i - f_2(\boldsymbol{X})\|_2 - \gamma\,]_+}}{\sum_{j=1}^{|\mathcal{E}|} \exp^{-[\,\|\mathcal{P}_j - f_2(\boldsymbol{X})\|_2 - \gamma\,]_+}}$
  - each event class → an energy-based hypersphere
- Sentence-level Energy Function

$$E_\Theta(\boldsymbol{X}, \boldsymbol{Y}) = -\left(\sum_{i=1}^{|\mathcal{E}|} \underbrace{\boldsymbol{Y}_i \left(V_{2,i}^\top f_2(\boldsymbol{X}) \right)}_{local} + \underbrace{w_2^\top g(W_2 \boldsymbol{Y})}_{label} \right)$$

- Loss Function

$$\mathcal{L}_{sen} = \sum_{i=1}^{K} \left[\triangle \left(\tilde{\boldsymbol{Y}}_i, \boldsymbol{Y}_i \right) - E_\Theta \left(\boldsymbol{X}_i, \tilde{\boldsymbol{Y}}_i \right) + E_\Theta \left(\boldsymbol{X}_i, \boldsymbol{Y}_i \right) \right]_+ + \mu_2 \mathcal{L}_{\mathrm{CE}} \left(\tilde{\boldsymbol{Y}}_i, \boldsymbol{Y}_i \right)$$



Three-Level Energy for Three Tasks

❑ Event-Relation Extraction

❑ Document-level Energy Function

$$E_\Theta(\ddot{\boldsymbol{X}}, \boldsymbol{z}) = -\left(\sum_{i=1}^{|\mathcal{R}|} \underbrace{\boldsymbol{z}_i \left(V_{3,i}^\top f_3(\ddot{\boldsymbol{X}})\right)}_{local} + \underbrace{w_3^\top g(W_3 \boldsymbol{z})}_{label}\right)$$

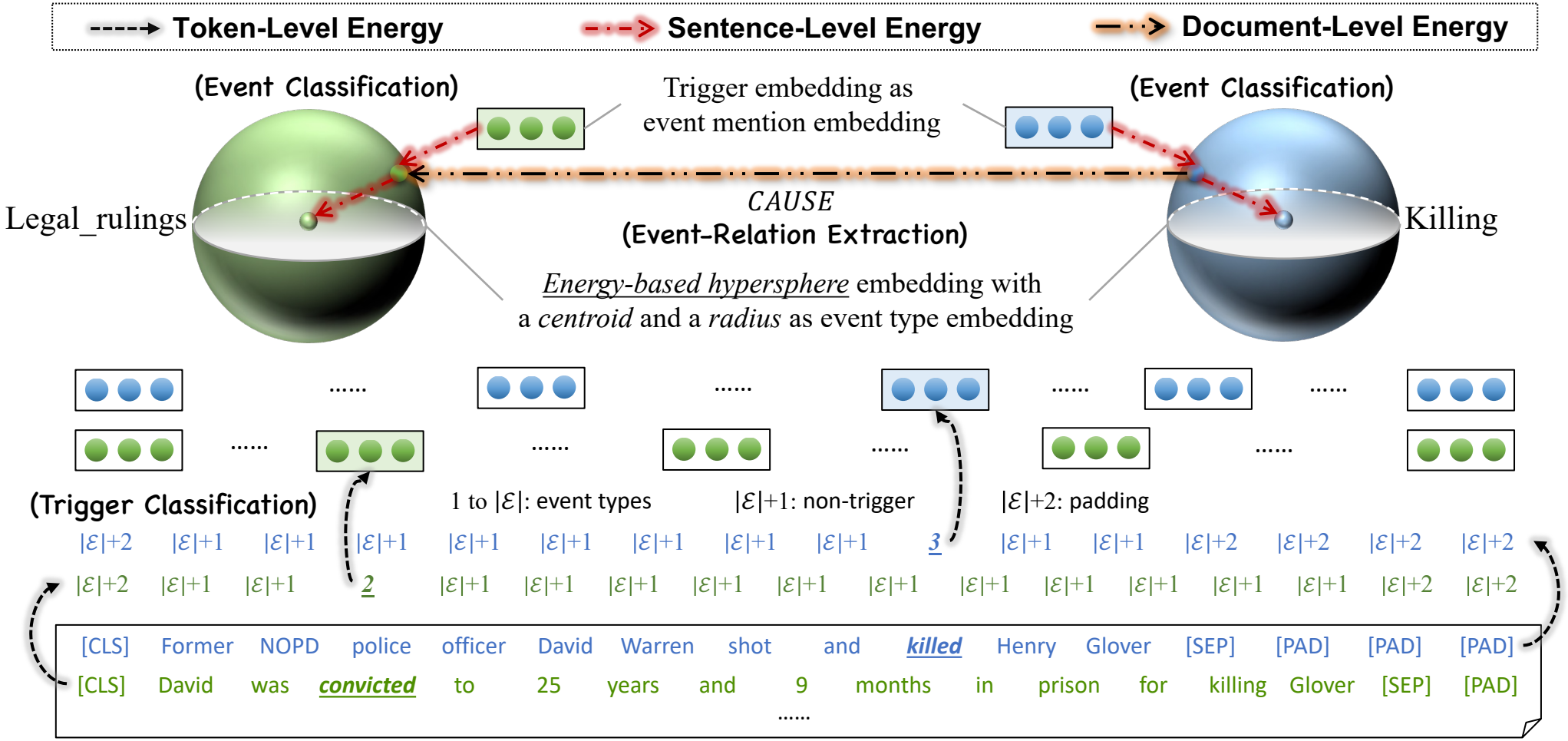
$$f_3(\ddot{\boldsymbol{X}}_{\langle ij \rangle}) = \left[f_2(\boldsymbol{X}_i), f_2(\boldsymbol{X}_j), f_2(\boldsymbol{X}_i) \odot f_2(\boldsymbol{X}_j)\right]$$

❑ Loss Function

$$\mathcal{L}_{doc} = \sum_{k=1}^{N} \left[\triangle\left(\tilde{\boldsymbol{z}}_k, \boldsymbol{z}_k\right) - E_\Theta\left(\ddot{\boldsymbol{X}}_k, \tilde{\boldsymbol{z}}_k\right) + E_\Theta\left(\ddot{\boldsymbol{X}}_k, \boldsymbol{z}_k\right)\right]_+ + \mu_3 \mathcal{L}_{\mathrm{CE}}\left(\tilde{\boldsymbol{z}}_k, \boldsymbol{z}_k\right)$$

Three-Level Energy for Three Tasks

$$\mathcal{L} = \lambda_1 \mathcal{L}_{tok} + \lambda_2 \mathcal{L}_{sen} + \lambda_3 \mathcal{L}_{doc} + \|\Phi\|_2^2$$



CONTENTS

Datasets and Baselines

❑ Datasets

Requirements: fine-grained annotations for events, such as labels of tokens, event mentions, and event-relations

- ❑ MAVEN-ERE
- ❑ OntoEvent-Doc

	MAVEN-ERE	ONTOEVENT-DOC
# Document	4,480	4,115
# Mention	112,276	60,546
# Temporal	1,216,217	5,914
# Causal	57,992	14,155
# Subevent	15,841	/

❑ Baselines

- ❑ Trigger Classification:
  - ❑ DMCNN, DMBERT, BiLSTM-CRF, BERT-CRF, MLBiNet, CorED-BERT, TANL, TEXT2EVENT
- ❑ Event Classification:
  - ❑ DMCNN, DMBERT, Hyperspherical Prototype Network (HPN), OntoED, TANL, TEXT2EVENT
- ❑ Event-Relation Extraction:
  - ❑ RoBERTa

OntoED: Low-resource Event Detection with Ontology Embedding (ACL 2021)
MAVEN-ERE: A Unified Large-scale Dataset for Event Coreference, Temporal, Causal, and Subevent Relation Extraction (EMNLP 2022)

Trigger Classification

consider cross-sentence semantic information

incorporate type-level and instance-level correlations

Model	MAVEN-ERE			ONTOEVENT-DOC		
	P	R	F1	P	R	F1
DMCNN†	60.09 ± 0.36	60.34 ± 0.45	60.21 ± 0.21	50.42 ± 0.99	52.24 ± 0.46	51.31 ± 0.39
BiLSTM-CRF†	61.30 ± 1.07	64.95 ± 1.03	63.06 ± 0.23	48.86 ± 0.81	55.91 ± 0.56	52.10 ± 0.43
DMBERT†	56.79 ± 0.54	<u>76.24</u> ± 0.26	65.09 ± 0.32	53.82 ± 1.01	<u>66.12</u> ± 1.02	59.32 ± 0.24
BERT-CRF†	62.79 ± 0.34	70.51 ± 0.94	65.73 ± 0.57	52.18 ± 0.81	62.31 ± 0.45	56.80 ± 0.53
MLBiNet‡	63.50 ± 0.57	63.80 ± 0.47	63.60 ± 0.52	56.09 ± 0.93	57.67 ± 0.81	56.87 ± 0.87
TANL‡	<u>68.66</u> ± 0.18	63.79 ± 0.19	66.13 ± 0.15	57.73 ± 0.65	59.93 ± 0.31	59.13 ± 0.52
TEXT2EVENT‡	59.91 ± 0.83	64.62 ± 0.65	62.16 ± 0.25	52.93 ± 0.94	62.27 ± 0.49	57.22 ± 0.75
CorED-BERT‡	67.62 ± 1.03	69.49 ± 0.63	<u>68.49</u> ± 0.42	<u>60.27</u> ± 0.55	62.25 ± 0.66	<u>61.25</u> ± 0.19
SPEECH	**78.82** ± 0.82	**79.37** ± 0.75	**79.09** ± 0.82	**74.67** ± 0.58	**74.73** ± 0.62	**74.70** ± 0.58
w/o energy	76.12 ± 0.32	76.66 ± 0.25	76.38 ± 0.28	71.76 ± 0.38	72.17 ± 0.39	71.96 ± 0.38

Table 2: Performance (%) of trigger classification on MAVEN-ERE *valid set* and ONTOEVENT-DOC *test set*. †: results are produced with codes referred to Wang et al. (2020b); ‡: results are produced with official implementation. **Best results** are marked in bold, and the <u>second best results</u> are underlined.

Event Classification

Model	MAVEN-ERE			ONTOEVENT-DOC		
	P	R	F1	P	R	F1
DMCNN	61.74 ± 0.32	63.11 ± 0.34	62.42 ± 0.15	51.52 ± 0.87	52.84 ± 0.61	52.02 ± 0.36
DMBERT	59.45 ± 0.48	**77.77** ± 0.21	67.39 ± 0.25	57.06 ± 1.04	**72.97** ± 1.11	**65.03** ± 0.45
HPN	62.80 ± 0.72	62.62 ± 0.99	62.71 ± 0.85	61.18 ± 0.81	60.88 ± 0.79	61.03 ± 0.81
OntoED	67.82 ± 1.70	67.72 ± 1.52	67.77 ± 1.61	**64.32** ± 1.15	64.16 ± 1.31	64.25 ± 1.22
TANL	68.73 ± 0.16	65.65 ± 0.63	67.15 ± 0.29	60.34 ± 0.71	62.52 ± 0.43	61.42 ± 0.51
TEXT2EVENT	61.14 ± 0.80	65.93 ± 0.69	63.44 ± 0.19	56.76 ± 0.97	66.78 ± 0.48	61.36 ± 0.77
SPEECH	**72.91** ± 0.76	72.81 ± 0.76	**72.86** ± 0.77	58.92 ± 0.96	58.45 ± 1.08	58.69 ± 1.40
w/o energy	71.22 ± 0.58	71.07 ± 0.45	71.12 ± 0.45	56.12 ± 1.87	55.69 ± 1.66	55.91 ± 1.76

Table 3: Performance (%) of event classification on MAVEN-ERE *valid set* and ONTOEVENT-DOC *test set*.

Event-Relation Extraction

ERE Task		RoBERTa	SPEECH
Temporal	MAVEN-ERE	**49.21** ± 0.33	39.64 ± 0.79
	+joint	**49.91** ± 0.58	40.23 ± 0.34
	ONTOEVENT-DOC	37.68 ± 0.47	**52.36** ± 0.71
	+joint	35.63 ± 0.70	**65.69** ± 0.39
Causal	MAVEN-ERE	**29.91** ± 0.34	16.28 ± 0.53
	+joint	**29.03** ± 0.91	16.31 ± 0.97
	ONTOEVENT-DOC	35.48 ± 1.77	**79.29** ± 2.15
	+joint	44.99 ± 0.29	**67.76** ± 1.28
Subevent	MAVEN-ERE	19.80 ± 0.44	**19.91** ± 0.52
	+joint	19.14 ± 2.81	**21.96** ± 1.24
All Joint	MAVEN-ERE	34.79 ± 1.13	**37.85** ± 0.72
	ONTOEVENT-DOC	28.60 ± 0.13	**54.19** ± 2.28

Table 4: F1 (%) performance of ERE on MAVEN-ERE *valid set* and ONTOEVENT-DOC *test set*. "+joint" in the 2_{nd} column denotes jointly training on all ERE tasks and evaluating on the specific one, with the same setting as Wang et al. (2022). "All Joint" in the last two rows denotes treating all ERE tasks as one task.

CONTENTS

Take Away

❑ Summary

- ❑ We propose a SPEECH model to tackle event-centric structured prediction with energy-based hyperspheres
- ❑ Experimental results demonstrate that SPEECH model is able to model manifold event structures with dependency and obtain effective event representations

❑ Outlook

- ❑ To model more complicated structures
- ❑ To test on more structured prediction tasks
- ❑ Maybe we can also equip LLMs with energy functions?

61 ACL 2023

NUS National University of Singapore

ZHEJIANG UNIVERSITY

Thank You

Speaker: Shumin Deng

Date: July, 2023

Preprint

Project

Preprint: http://arxiv.org/abs/2305.13617, Project: https://github.com/zjunlp/SPEECH

Toronto, Canada

July 9-14, 2023