Title
Towards Universal Event Extraction

ISBN
9798247909392

Author
Parekh, Tanmay Devendra

Publication Date
2026-05-22

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Towards Universal Event Extraction

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Computer Science

by

Tanmay Devendra Parekh

2026

ABSTRACT OF THE DISSERTATION

Towards Universal Event Extraction

by

Tanmay Devendra Parekh

Doctor of Philosophy in Computer Science

University of California, Los Angeles, 2026

Professor Nanyun Peng, Co-Chair

Professor Kai-Wei Chang, Co-Chair

Extracting structured knowledge from unstructured text "in the wild" remains a fundamental challenge for AI, often bottlenecked by data scarcity and the brittle reasoning of standard models in specialized domains. This dissertation introduces a unified framework for Universal Event Extraction that successfully bridges the gap between open-domain generalizability and high-stakes real-world applications. To drive this generalizability, we propose targeted synthetic data generation pipelines that dynamically adapt to novel schemas and domains, overcoming the manual annotation bottleneck. Simultaneously, to resolve the discriminative reasoning failures of off-the-shelf language models, we architect specialized multi-agent frameworks that enhance planning and reasoning by decomposing complex extraction into iterative, verifiable steps. Crucially, this combined approach can now extract arbitrary event types across new languages and domains in completely zero-shot and low-resource settings. Grounding the application of this universal paradigm, we demonstrate its profound real-world efficacy by successfully deploying it for automated global epidemic forecasting. Ultimately, this thesis establishes a scalable foundation for transforming raw text into structured event graphs, powering deeper semantic understanding and strategic decision-making for next-generation AI models.

The dissertation of Tanmay Devendra Parekh is approved.

Wei Wang

Aditya Grover

Kai-Wei Chang, Committee Co-Chair

Nanyun Peng, Committee Co-Chair

University of California, Los Angeles

2026

To my wife, my parents, my extended family, and my friends – for their unwavering trust, love and support.

Contents

List of Figures

List of Tables

ACKNOWLEDGEMENTS

Throughout my professional and personal life, I have been exceptionally fortunate to cross paths with many wonderful individuals. These interactions have profoundly shaped who I am today, and I want to take this opportunity to express my deepest gratitude for their unwavering support, encouragement, and guidance. This backing has been truly invaluable; simply put, I would not be where I am without it.

First and foremost, I express my deepest gratitude to my advisors, Prof. Kai-Wei Chang and Prof. Nanyun Peng. When I first met them, I was unaware they were married, but throughout this journey, they have truly made me feel like their academic child. They nurtured my professional growth and maintained steadfast confidence in my capabilities, even when my own self-trust wavered. During the early, uncertain months of my PhD, as I explored various research topics without much success, Violet generously dedicated her time to discuss low-level experimental details with me. Her patience directly enabled me to complete my first project after a year and a half. Concurrently, Kai-Wei empowered me to explore broader research directions, constantly instilling the belief that I had what it takes to succeed as a doctoral student. Together, they provided ample opportunities for me to develop my leadership, mentorship, and social skills through roles such as social chair, SoCal NLP PC, and the UCLA NLP Seminar committee. Before choosing a doctoral program, a senior PhD student, Wasi, told me that beyond their professional excellence, both Kai-Wei and Violet are simply wonderful human beings. That was the reason I chose UCLA. As I conclude my journey here, I could not agree more. I am filled with immense praise and gratitude for their extraordinary support in both professional and personal capacities.

I would also like to thank my committee members, Prof. Wei Wang and Prof. Aditya Grover, for their valuable time, logistical guidance, and constructive feedback throughout the completion of my degree. On a personal note, I had the distinct fortune of working with Prof. Wei on the multi-year SPEED/SPEED++ project. I am deeply grateful for her continuous support and unwavering belief in my ability to lead that initiative. It stands as one of the

proudest achievements of my PhD, and I sincerely thank her for the opportunity.

I extend my sincere appreciation to the organizations that funded my research. I thank the UCLA Computer Science Department for providing me with a fellowship and Teaching Assistant opportunities in my initial years. I am also grateful to the National Science Foundation (NSF) and Optum Labs for supporting my mid-year graduate research, as well as Amazon and Bloomberg for granting me the Science Hub Fellowship and the Data Science PhD Fellowship for my final years. This financial support was instrumental in smoothing the path of my PhD. Equally important are the UCLA Computer Science administrative staff, whose hard work minimized my logistical distractions. A special shout-out to Helen for managing PhD logistics and registration; Juliana for handling course information and stepping in during Helen's absence; Chris, Edna, and Osanna for processing reimbursement requests; Joann for her months of effort securing my fellowship funds; and the CS IT team for managing server requests. I also thank the Dashew Center for their diligent assistance with my visa status. Finally, I appreciate UCLA Housing for an amazing three-year stay at the Rose Apartments, and the Rose community for helping me build lasting friendships and expand my network.

To my lab and all my labmates: thank you for cultivating such a vibrant environment that fostered my passion for research. I will always cherish the memories from our trips to Santa Barbara, San Diego, Arrowhead, and Irvine, as well as conference travels to Abu Dhabi, Toronto, Miami, and Mexico City. I am grateful to Kai-Wei and Violet for funding these excursions and cultivating a strong sense of community. I must take a moment to single out I-Hung Hsu and Kuan-Hao Huang, my senior PhD mentors and close collaborators. You both invested heavily in me during my first three years, and I owe a significant part of my professional success to your mentorship. Beyond our weekly meetings, they dedicated countless extra hours to nurture my skills, offer professional advice, aid my networking, and build my self-confidence. Thank you, Yufei Tian, not only as a senior PhD but as a remarkable friend. When I was anxious since I did not publish in my second year, she took the time to

share your own experiences and motivate me; she has been a steadfast cheerleader ever since. Thank you to senior members Jia-Chen, Alex, and Sidi for your profound influence on my journey, and to current members Hritik, Di, Bryan, Xueqing, Po-Nien, Lucas, Haikang, Elaine, Rohan, and Mohsen for years of deep, insightful research discussions. I appreciate Christina, Haoyi, and Yufei for their help with social chair duties, and Lucas, Mohsen, Bryan, Haikang, and Wenbo for the engaging outside-work conversations during group lunches. Ultimately, an overarching thank you to everyone in Kai-Wei's lab (Pan, Kareem, Kuan-Hao, Wasi, Jieyu, Arjun, Fan, Wade, Elia, Masoud, Harold, Tao, Xiao, Christina, Joey, Amita, Elaine, Rui, Di, Hritik, Xueqing, Bryan, Wenbo, Eric, Steven) and Violet's lab (Sarik, RJ, Te-Lin, I-Hung, Sidi, Zi-Yi, Chujie, Yufei, Jiao, Alex, Tuhin, Fabrice, Po-Nien, Lucas, Mohsen, Haoyi, Haikang). You have all played vital roles in this journey.

I am equally grateful to all my collaborators. Beyond my advisors and senior mentors I-Hung and Kuan-Hao, I extend my appreciation to the PhD and Master's students I worked alongside - Ashima, Xiao, Lucas, and Hyosang - from whom I learned a great deal. I also offer a huge thank you to the brilliant undergraduate students I had the privilege to mentor and grow with: Alex, Jiarui, Yuxoan, Syed, Bonnie, Eric, Jeffrey, Sparsh, Sreya, and Artin. During my PhD, I had the opportunity to build invaluable industry connections. From my time at Meta, I thank Denis for being an exceptional mentor who taught me to shut out the noise and focus on personal research goals. I thank my undergrad senior Pradyot for his steadfast support, and Akshay and Alex for their guidance during the internship. At Amazon, Kartik and Ninareh were phenomenal sounding boards for my fellowship project, providing industry perspectives that continue to shape my career decisions. At Bloomberg, I am indebted to Yunmo, a close collaborator and mentor whose insights have been consistently profound, as well as Ella, Shuyi, Sachith, and Srivas for their extensive support. I also thank the Optum Labs team - Sanjit, Dongxu, Ehsan, Masoud, Zhichao, and Robert - for a wonderful year of weekly collaborations. Finally, to the team members of ongoing or exploratory collaborations - Steven, Zihan, Yufei, Alex, Aryan, Yu-Chi, Anshul, Chris, Rebecca, and Aman - even without

official publications yet, the shared experiences have been immense learning opportunities.

Reflecting on my Master's at Carnegie Mellon University (CMU), several individuals laid the groundwork for my PhD. I am deeply thankful to my advisor, Prof. Alan Black, and collaborating advisors, Prof. Yulia Tsvetkov and Prof. Graham Neubig, for providing the research opportunities and guidance that sparked my academic growth. I also thank Prof. Alex Rudnicky for his funding and partnership on the Alexa Socialbot Challenge. To my collaborators and friends Amrith Setlur and Aman Madaan: thank you for an incredible two years of learning and weathering the ups and downs together. I appreciate senior PhDs - Shrimai and Emily - for their incredible kindness and patience in mentoring me long after our formal collaborations ended. Lastly, thank you to Rishabh, Sopan, Aditya, Sai, Sanket, Aditi, and Prakhar for the countless enriching discussions at CMU.

Tracing back to the very origins of my research journey, I must thank my mentors from my undergraduate years. Prof. Preethi Jyothi's Automatic Speech Recognition (ASR) course initially inspired my transition into Natural Language Processing. Her year and a half of mentorship led to my first publications; she predicted I would pursue a PhD, and though I doubted it then, I could not have been more wrong. I am also deeply grateful to Saurabh Garg, a batchmate and close collaborator during those formative years; we authored early papers together, and I am thrilled we still push each other to greater heights. Prof. Shivaram Kalyanakrishnan provided invaluable alternative research avenues and strong recommendations for my studies abroad. I also wish to honor the late Prof. Pushpin Bhattacharya and Tejas Srinivasan for fueling my early passion for research. Finally, I thank my early industry mentors: Sunayana and Monojit from Microsoft Research India, alongside Sachin, Arijit, Ram, Ayush, and Shweta from the Amazon Machine Learning team.

Stepping away from professional ties, I want to celebrate the friends who have shaped my life. To my high school friends - Mayank, Sambhav, Divyesh, and Meet - thank you for 15-20 years of unwavering friendship. I appreciate my IIT coaching friends, including Sanket, Dharmin, Swapnil, Kaustabh, Rohit, and Jyesht, many of whom remain close today. To my

undergraduate core from IIT Bombay - Ritwick, Pratyarth, and Navneet - thank you for being my anchors for those four years. I also cherish the memories with my broader IIT circle: Pranit, Pranav, Anmol, Chinmay, Aditya, Adithya, Mihir, Hardik, Aviral, Saurabh, Siddhant, Sanat, Shachi, Shievani, and others. To my CMU family - Ritwick, Amala, Soumya, Bhuvan, and Rohit - thank you for making my time there so memorable and for carrying me through one of my most difficult phases. And to the friends who defined my PhD years - Aarjav, Dipti, Nischal, Harsh, Akash, Amrit, Manoj, Ayush, Swarali, Poorva, Pragya, Anwesha, Niharika, Shruti, Hritik, Ashima, Aditya, Hemil, Chinmay, and Ashutosh - thank you for filling these years with dinners, walks, festivities, dance parties, hikes, and travels.

Next, my deepest gratitude belongs to my family. Thank you to my relatives across the US - Chintu mama, Mukesh uncle, Ankit bhaiya, Aditya bhaiya, Chinki masi, Chandresh chacha, Sanmukh bhaiya, Pranay mama, Mayank bhaiya, Devang, and your families - for welcoming me with open arms and making the States feel like home. Coming from a joint family, I have been blessed by the support of hundreds of relatives from both my maternal and paternal sides; your collective encouragement has been invaluable. I thank my grandparents for their constant, unconditional love and for instilling my core values. Over the course of this PhD, I was blessed to celebrate a beautiful wedding, and I want to extend a heartfelt thank you to my new family of in-laws for welcoming me so warmly. To my younger brother, Samarth: thank you for being such a central part of my life and for putting up with all my antics. I genuinely would not be here without you. To my parents, my ultimate pillars of support: thank you for your boundless trust, for encouraging me to dream bigger, and for celebrating every milestone with me.

Words cannot fully capture my love and gratitude for my partner, Dimple. She has been my pillar over the last few years, my light of hope, and my greatest source of happiness through the different seasons, through the thick and thin, through all my ups and downs. Sharing this journey with a fellow researcher who truly understands the unique rigors of a doctoral pursuit has been an immense privilege. I am incredibly lucky to have her by my

side as my partner for life.x

To everyone named here, and to anyone I may have inadvertently missed - a massive thank you. Every interaction has left an indelible mark, and I am profoundly grateful for your role in my story.

VITA

Education

Ph.D. in Computer Science	2021 - 2026
University of California Los Angeles	(Expected)
M.S in Computer Science	2019 - 2021
Carnegie Mellon University	
B.Tech. in Computer Science	2014 - 2018
Indian Institute of Technology Bombay	

Internships

Data Science Intern, Bloomberg AI	Jun '25 - Oct '25
Research Scientist Intern, Meta GenAI	Jun '24 - Oct '24
Applied Scientist Intern, Amazon Alexa	Jun '22 - Sep '22
Applied Scientist, Amazon	Jul '18 - Jun '19

Awards

Bloomberg Ph.D. Data Science Fellowship	2025 - 2026
Amazon Science Ph.D. Fellowship	2024 - 2025
UCLA Computer Science Fellowhsip	2021 - 2022
Outstanding Reviewer, EMNLP	2025
Best Paper Nomination, NAACL	2024
Program Chair, Socal NLP Symposium	2023
ISCA Student Grant, Interspeech	2018

SELECTED PUBLICATIONS

- **Tanmay Parekh**, E Hofmann-Coyle, S Wang, SSR Kothur, S Prasad, Y Chen. PExA: Parallel Exploration Agent for Complex Text-to-SQL. In Proceedings of the ACL 2026.

- **Tanmay Parekh**, K Mehta, N Mehrabi, KW Chang, and N Peng. DiCoRe: Enhancing Zero-shot Event Detection via Divergent-Convergent LLM Reasoning. In Proceedings EMNLP 2025.

- **Tanmay Parekh**, Y Dong, L Bandarkar, A Kim, IH Hsu, KW Chang, and N Peng. SNaRe: Domain-aware Data Generation for Low-Resource Event Detection. In Proceedings of EMNLP 2025.

- **Tanmay Parekh**, P Prakash, A Radovic, A Shekher, and D Savenkov. Dynamic strategy planning for efficient question answering with large language models. In Proceedings of the Findings of the NAACL 2025.

- **Tanmay Parekh**, J Kwan, J Yu, S Johri, H Ahn, S Muppalla, KW Chang, W Wang, and N Peng. SPEED++: A multilingual event extraction framework for epidemic prediction and preparedness. In Proceedings of EMNLP 2024.

- KH Huang, IH Hsu, **Tanmay Parekh**, Z Xie, Z Zhang, P Natarajan, KW Chang, N Peng, and H Ji. TextEE: Benchmark, Reevaluation, Reflections, and Future Challenges in Event Extraction. In Findings of the ACL 2024.

- **Tanmay Parekh**, IH Hsu, KH Huang, KW Chang, and N Peng. Contextual label projection for cross-lingual structured prediction. In Proceedings of the NAACL 2024.

- **Tanmay Parekh**, A Mac, J Yu, Y Dong, S Shahriar, B Liu, E Yang, KH Huang, W Wang, N Peng, and KW Chang. Event detection from social media for epidemic prediction. In Proceedings of NAACL 2024.

- **Tanmay Parekh**, IH Hsu, KH Huang, KW Chang, and N Peng. GENEVA: Benchmarking generalizability for event argument extraction with hundreds of event types and argument roles. In Proceedings of ACL 2023.

Chapter 1

Introduction

In an era characterized by unprecedented unstructured textual data, it is imperative to rapidly and accurately distill complex narratives into actionable, machine-readable knowledge. Event Extraction (EE) serves as this critical computational bridge, transforming raw text into structured representations of "who did what to whom, when, and where" [37]. This transformation has directly improved various downstream systems across various disciplines, ranging from automated financial forecasting [254], biomedical reasoning [114], knowledge graph construction [179] to real-time epidemiological surveillance [164, 163].

Traditional EE frameworks remain fundamentally under-utilized owing to their rigidity. These frameworks, developed on predefined limited event ontologies, fail to generalize to the long-tail of dynamic, real-world event occurrences [73]. This severely limits the utility of EE to broader range of real-world specialized tasks and domains. Recently, Large Language Models (LLMs) have offered stronger semantic understanding and generative capabilities that challenge these historical generalization barriers. However, even LLMs perform poorly on EE out of the box [86], owing to the extensive domain knowledge and the strong discriminative reasoning that the task Event Extraction demands (Chapter 2).

To this end, my thesis champions the paradigm of *Universal Event Extraction* – a vision that entails the development of robust, real-world adaptable frameworks capable of

extracting structured information for any event type, from any domain, and across any language. Realizing this vision requires advancing two distinct tracks: **Generalizability** and **Applications**. First, the Generalizability track demands enhancing models and datasets to maintain broad coverage across unseen event types, diverse text domains, and low-resource languages. To drive this generalizability, I leverage synthetic data generation to resolve persistent data scarcity, alongside multi-agent frameworks that inject the strategic reasoning necessary to improve the extraction capabilities of off-the-shelf LLMs. Second, the Applications track requires bridging the gap between these theoretical extraction capabilities and the diverse, practical demands of downstream tasks.

First, this thesis discusses the task of Event Extraction and provides a comprehensive background of the current research landscape in Chapter 2. Subsequently, I discuss works tackling these distinct challenges:

- **GENEVA (Chapter 3):** This work, focused on improving ontology generalization, introduced the largest event ontology and EE dataset in terms of argument role diversity by repurposing existing Semantic Role Labeling (SRL) data. This work pioneers the challenge of non-entity-based argument roles, significantly expanding the summarization capacity of EE.

- **CLaP (Chapter 4):** Addressing cross-lingual generalizability, in this work, we propose a novel technique of contextual machine translation using instruction-tuned language models for structured label projection. This approach yields highly accurate label translations and improves downstream performance across 47 languages, including 10 extremely low-resource languages.

- **DiCoRe (Chapter 5):** This work introduces a divergent-convergent multi-agent reasoning framework designed to elicit more accurate and comprehensive event structures from LLMs. The core mechanism to improve LLMs' reasoning is achieved by unconstraining the task in multiple stages.

- **SNaRe (Chapter 6):** To address data scarcity and reduce expert annotation costs, we propose a multi-agent data generation pipeline. Specifically, this work introduces domain-specific signals to the paradigm of inverse data generation to synthesize high-quality, complex training data, that can be utilized to train downstream models. Overall, this work paves the way for scalable EE with limited manual annotation.

- **SPEED/SPEED++ (Chapter 7):** This work, comprising two sub-works, is focused on demonstrating a strong application of EE for epidemiology. Specifically, we developed the first social media multilingual EE framework for monitoring and detecting upcoming epidemics. This involved creating a disease-agnostic epidemiological event ontology, corresponding EE datasets, and developing zero-shot cross-lingual EE models. On the application front, we utilized our models for cross-disease, cross-language epidemic forecasting and epidemic monitoring for public health tasks using real-world social media data.

Finally, I outline some existing open challenges and propose future directions necessary to fully realize the vision of Universal Event Extraction in Chapter 8.

Chapter 2

Background and Related Works

This thesis is primarily focused on the task Event Extraction. In this chapter, we provide a formal background of the task and discuss related works in this field that lay the foundation for our proposed work later.

2.1 Task Definition: Event Extraction

For the majority of our works, we follow the technical definitions for the task of event extraction as provided by Doddington et al. [37] (except for Chapter 3 where we provide the adjusted definitions in Section 3.2.1). We describe the important terminologies for this task along with some examples below.

An **event** is loosely defined as something that happens. Alternatively, it's also described as a specific occurrence that involves some participants. For our works, we limit the scope of events to a sentence or a few sentences; in contrast to other works [65, 121, 212] which discuss events across an entire document. Each event is associated with a specific **event name**, which is a semantically understood name for the event. The event name is also associated with an **event definition**, which elaborates the underlying meaning of the event name. Typically, the possible set of event names and definitions are pre-defined as part of the **event ontology**. An example event ontology comprising three event types *Attack*, *Sentence*,

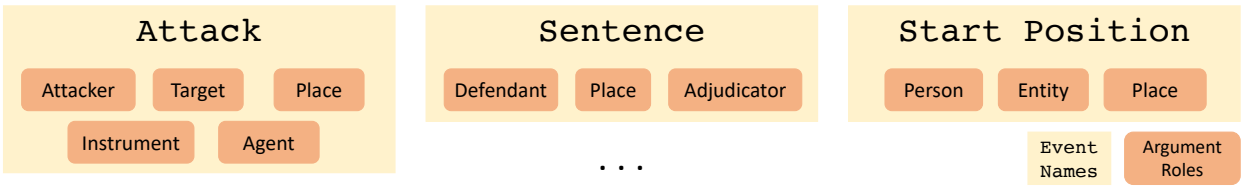

Figure 2.1: Example event ontology derived as a subset of the ACE [37] event ontology. The yellow boxes indicate the event names and the orange boxes indicate the corresponding argument roles (as shown in the legend in the bottom right). Event and argument role definitions are omitted in this figure.

and *Start Position*, is shown in Figure 2.1 which is a subset of the ACE [37] event ontology. An **event mention** is defined as the natural language sentence(s) in which the event is mentioned/described. By nature, there can be sentences with no events; while a single event mention can have multiple events in it. To distinguish events in the sentence, each event is also associated with an **event trigger**. Event triggers are typically single-word, but rarely span multiple words as well. An example event mention with the corresponding trigger word and event type is shown in Figure 2.2.

Event arguments are defined as the participants that provide salient and specific information/components of the event. An event mention can comprise zero to multiple arguments. Each argument is also associated with a specific **argument role**, representing the semantic category of the information the argument describes. Each argument role is also accompanied by an **argument role definition**, which elaborates on the semantic meaning of the argument role. Both argument roles and definitions are also pre-defined in the **event ontology**. Argument roles are a unique characteristic for each event type; thus each event type can have different sets of argument roles, as shown in Figure 2.1. Simultaneously, different event types can share the same argument role as well (e.g. *Place* argument role is shared for all three event types in the event ontology in Figure 2.1). Similar to event mentions, an **argument mention** is the natural language sentence(s) where the argument is mentioned. Arguments are typically entities in the argument mentions, but we also explore non-entity arguments in Chapter 3. An example of arguments and their corresponding roles

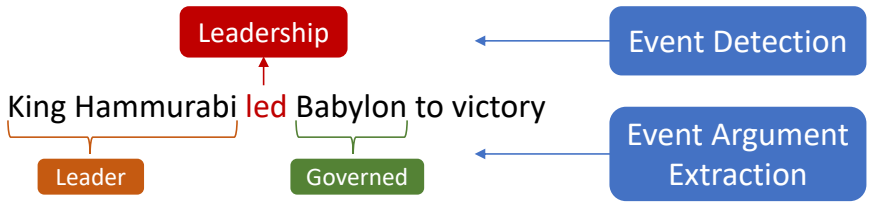

Figure 2.2: Illustration of the task of Event Extraction (EE). Event Detection (ED) extracts the trigger word (*led*) and classifies into the event type *Leadership*. Event Argument Extraction (EAE) uses this information to extract the arguments (*King Hammurabi* and *Babylon*) and classify them into their respective roles (*Leader* and *Governed*). The joint extraction of the events and the arguments is the task of Event Extraction.

is shown in Figure 2.2.

The overall task of **Event Extraction (EE)** is usually split as two subtasks:

1. The first subtask of **Event Detection (ED)** aims to identify the event triggers from the sentence and classify them into the pre-defined set of event types (as defined in the event ontology). This subtask is typically evaluated using F1 scores for *Trigger Identification* (Tri-I) i.e. identification of event triggers, and *Trigger Classification* (Tri-C) i.e. classification of identified triggers into the pre-defined event types [7].

2. The second subtask of **Event Argument Extraction (EAE)** utilizes the given event in the sentence to extract the arguments corresponding to this event from the sentence and classify them into the pre-defined event-specific argument roles. Note that the event trigger and corresponding event type are provided as inputs for this task. This task is evaluated using F1 scores for *Argument Identification* (Arg-I) i.e. identification of arguments, and *Trigger Classification* (Arg-C) i.e. classification of identified arguments into the pre-defined roles [7].

The joint extraction of the events and the corresponding arguments from the sentence is the motive of the task of **Event Extraction (EE)**. Naturally, it is evaluated using the F1 scores for Tri-I, Tri-C, Arg-I, and Arg-C. An illustration of this task along with the two subtasks is shown in Figure 2.2.

2.2 Background and Limitations: Datasets and Models

In this section, we discuss the previous works for creating EE datasets and models, along with their limitations with respect to generalizability and applicability.

Event Extraction Datasets: The two most commonly used datasets for Event Extraction include ACE05 [37] and RichERE [196] (also known as ERE) - both of which focus on the news domain covering a limited number of 33 event types and 22 argument roles. These datasets are also 10-20 years old and are distributionally divergent to the current data trends. Later, RAMS [45] neutralized this distributional gap along with wider range of event types, but they as well focus on the news domain. MAVEN [220] introduced a large ED dataset with 168 event types in the general domain. FewEvent [33], another general domain ED dataset, introduces the few-shot setting for ED. Next, WikiEvents [121] introduced a large document-level dataset in the Wikipedia domain MEE [171] was another dataset in the Wikipedia domain focusing on limited event types, but introducing multilingual annotations in eight languages. Recently, GLEN [122] introduced a 1000+ event type dataset for ED which is partially LLM-generated. M^2E^2 [117] and VOANews [119] also introduce event extraction in the multimodal domain. There have been some domain-specific datasets as well including MLEE [173], Genia2011 [95], and Genia2013 [96] from the biomedical domain; CASIE [184] from the cybersecurity domain; PHEE [203] from the pharmacovigilance domain.

In terms of generalizability, most datasets except RAMS, FewEvent, and MAVEN have a limited event type coverage. Furthermore, all of these datasets focus only on a smaller limited set of argument roles. This demonstrates the shortcoming of limited event coverage. In terms of domain coverage, there have been a recent spread of datasets across different domains, but the major modeling benchmarking is still done on the earlier datasets of ACE and ERE. The specialized in-domain are never used outside the application-specific models for benchmarking. This introduces domain specificity in various models, as is highlighted in Huang et al. [86]. From a multilingual perspective, only ACE [37] and MEE [171] have

non-English annotations, and cover only up to 10 unique languages in total, underlining the language diversity in EE.

Event Extraction Models: There are two major paradigms of EE modeling: (1) Pipelined and (2) End-to-end. Pipelined models train on the subtasks of ED and EAE independently and during inference, use these independent models in a pipelined fashion. End-to-end modeling combines both the annotations and trains a joint EE model. Some models are flexible to train/infer using both paradigms, but most models adhere to one of the two.

Following [86], we can organize the different EE models into three major categorizations: (1) *Joint training models*, the pure end-to-end fashion models, including models like DyGIE++ [216], OneIE [125], and AMR-IE [253]. (2) *Classification-based models*, that utilize a classification head over a base model to classify the different event types/argument roles. The base model can be a simple token embedding [82], a sequential labeling model [74, 122], a question-answering model [42, 218], or a machine comprehension model [127]. (3) *Generation-based models*, that utilize generative models to reformulate event extraction as conditional generation. These include models like BART-Gen [121], DEGREE [73], PAIE [135], and AMPERE [76].

In practical downstream applications, we majorly face new event types which may not be defined in the existing EE datasets. Most of the models are confined to restricted sets of event types, domains, and languages on which they are trained on - thus cannot generalize to wider and new set of event types. To introduce this ability to generalize to these new event types, various works work in the direction of zero-shot EE, which assumes no training on new event types, with limited/no training on a set of base event types. These works majorly include techniques utilizing transfer learning [87], word sense disambiguation (WSD) [241], modeling event definitions [251], or utilizing label semantics [42, 250, 133, 73]. But these works perform poorly at an absolute scale making them unusable for practical downstream applications. Furthermore, there are only limited works which focus on generalizing across

domains and languages [6, 85] and this area is largely unexplored.

Event Extraction in the era of LLMs: Recent years have witnessed the rise of Large Language Models (LLMs) like ChatGPT [18], Llama [213, 214], Mixtral [91], etc. which have practically solved many NLP tasks using their in-context and emergent reasoning capabilities [225, 174]. Several works [56, 86] have utilized these LLMs through various prompts and chat-styled approaches, but all of them conclude that there is a significant gap between the in-context LLM performance and fine-tuned models on this task. This indicates how the deeper semantic reasoning and the requirement of a structured output can be a challenging task even for large language models, as also noted in [112, 86].

Event Extraction and practical applications: Some works have explored the usage of EE for other related NLP tasks like temporal relation extraction [66], building knowledge graphs [249], question-answering [16], and bias detection [202]. Hogenboom et al. [72], Yadav et al. [234] survey EE for usage in decision support, while there have been various works in the clinical domains partially utilizing EE for drug discovery [219], adverse case detection [178], workflow optimizations [78]. Some works [108, 140] in the epidemiological domains for surveillance also utilize some concepts from EE. Overall, there has been some academic discussion and exploration on the applicability of EE for practical purposes, but this direction of research still does not utilize the complete potential of EE and remains largely unexplored with respect to broader disciplines like law, finance, policymaking, etc.

2.3 Is Event Extraction solved by LLMs?

As Large Language Models (LLMs) demonstrate remarkable proficiency across a spectrum of applications – from basic sentiment analysis to complex mathematical reasoning – a natural question arises: is the specialized study of Event Extraction (EE) still relevant, or have LLMs already "solved" this task? The empirical evidence indicates that, despite their broad

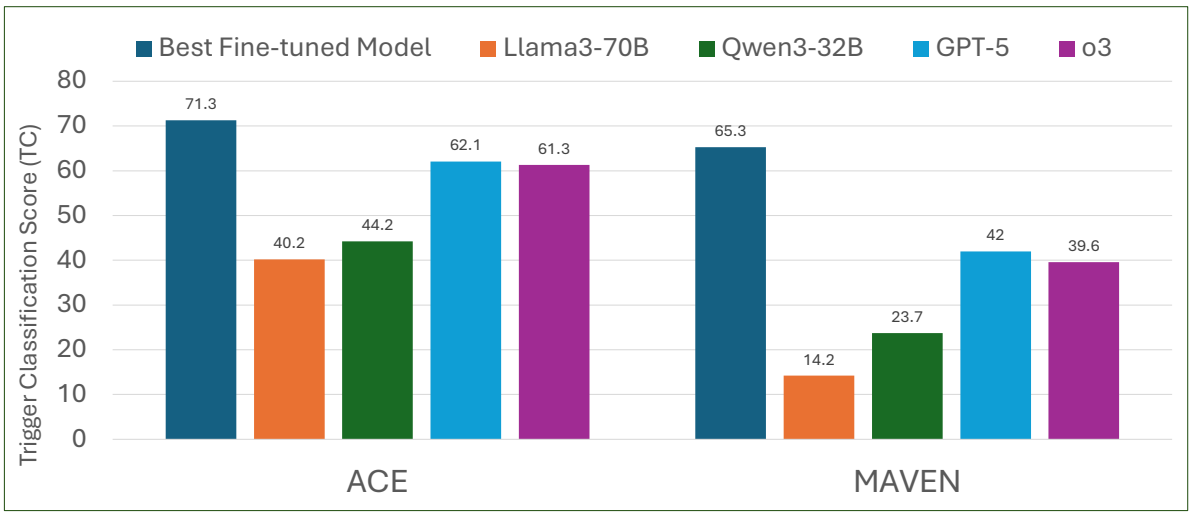

Figure 2.3: Comparison of best closed and open-source LLMs on Event Extraction relative to the best fine-tuned small LM framework.

capabilities, off-the-shelf LLMs remain notably deficient at structural event extraction.

Recent surveys [86] have documented the performance deficits of zero-shot LLMs when compared to fully supervised, smaller language models. To provide a fresh perspective on this reasoning gap, we evaluate recent, state-of-the-art LLMs against two foundational EE datasets: ACE 2005 [37] and MAVEN [220], as illustrated in Figure 2.3.

Measuring performance via the F1 score for event trigger classification reveals a stark contrast. On the widely established ACE dataset – where the risk of data contamination and memorization by closed-source LLMs is highest – the best fine-tuned, task-specific models still outperform state-of-the-art closed LLMs by 9 to 10%. This disparity drastically widens on the more recent and structurally challenging MAVEN dataset, where the performance gap increases to 22 to 25%. Furthermore, when evaluating open-weights LLMs, the deficits are even more pronounced, trailing supervised baselines by 25 to 40% across both datasets.

These findings underscore a critical finding: Event Extraction is not yet a solved problem in the era of generative AI. Addressing this discriminative reasoning gap is one of the primary motivation of this dissertation, which seeks to bridge the divide by architecting specialized multi-agent frameworks that harness the latent reasoning and generative power of LLMs for high-fidelity event extraction.

Chapter 3

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

Recent works have focused on building generalizable EAE models [87, 133, 182] and they utilize existing datasets like ACE [37] and ERE [196] for benchmarking. However, as shown in Figure 3.1, 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 GENEVA 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

[1] Abstract event types are defined as the top nodes of the event ontology created by MAVEN [220].

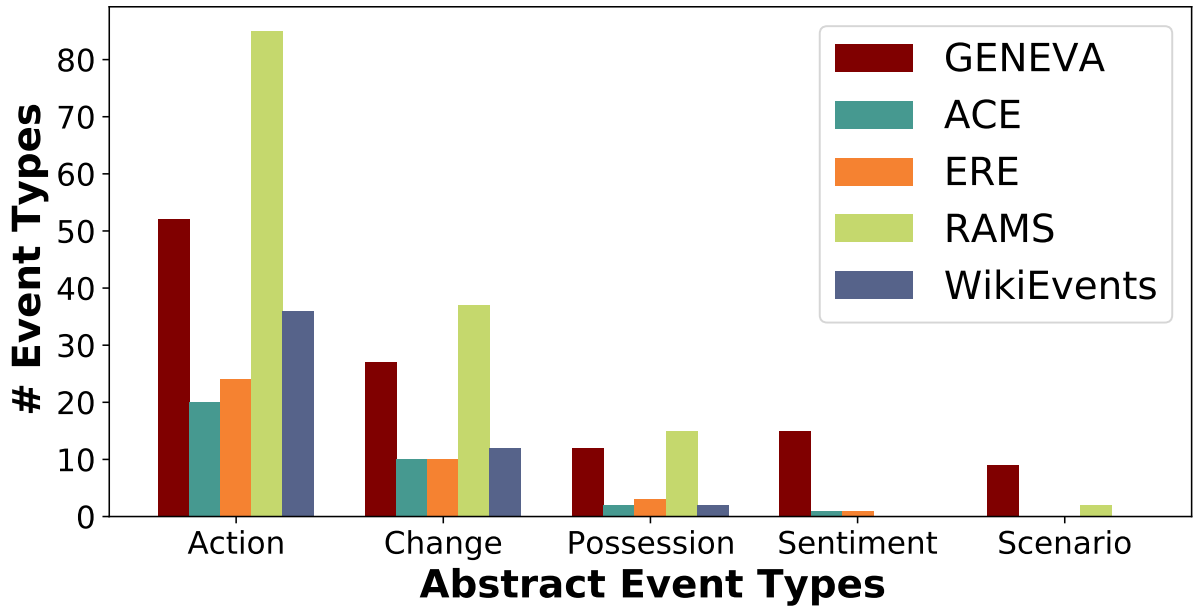

Figure 3.1: Distribution of event types into various abstract event types[1] for GENEVA, ACE [37], ERE [196], RAMS [45], and WikiEvents [121] datasets. We observe that GENEVA is relatively more diverse than the other datasets.

labeling (SRL) and EAE [3] and leverage a diverse and exhaustive SRL dataset, FrameNet [14], to build the ontology. Through extensive human expert annotations, we design mappings 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 [216, 125, 218], question-answering-based models [42], and genera-

tive approaches [159, 73]. We also introduce new automated refinements in the low resource state-of-the-art model DEGREE [73] 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.

Overall, in this chapter, our work makes the following contributions. (1) 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. (2) We propose a generalizability evaluation dataset GENEVA and benchmark various recent EAE models. (3) We show how GENEVA is a challenging dataset, thus, encouraging future research for generalization in EAE.

We first introduce some related works in Section 3.1. Then we discuss how we create our ontology based on the FrameNet dataset with detailed steps in Section 3.2. Next, we discuss how we create our dataset based on this ontology, provide some statistics and introduce our benchmarking setups in Section 3.3. We finally present the results of benchmarking various models on our dataset along with deeper analysis to show how our dataset can be challenging in Section 3.4. The work in the chapter has been published in Parekh et al. [161].

3.1 Background and Related Works

Since our work introduces a new dataset and benchmarks various models on it, we majorly discuss and recap some existing Event Extraction datasets and models in this section.

Event Extraction Datasets and Ontologies The earliest datasets in event extraction date back to MUC [204, 61]. Doddington et al. [37] 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 [196] and various TAC KBP Challenges [47, 48, 59].

These datasets cover a small and restricted set of event types and argument roles with limited diversity. Later, MAVEN [220] 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 [45], WikiEvents [121], and DocEE [212]; but their ontologies are also entity-centric, and their event coverage is limited to specific abstract event types (Figure 3.1). 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 [147]. Previous classification-based approaches have utilized pipelined approaches [238, 216] as well as incorporating global features for joint inference [120, 236, 125]. However, these approaches exhibit poor generalizability in the low-data setting [127, 73]. To improve generalizability, some works have explored better usage of label semantics by formulating EAE as a question-answering task [127, 116, 42]. Recent approaches have explored the use of natural language generative models for structured prediction to boost generalizability [186, 188, 159, 121]. Another set of works transfers knowledge from similar tasks like abstract meaning representation and semantic role labeling [87, 133, 250]. DEGREE [73] 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.

3.2 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,

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

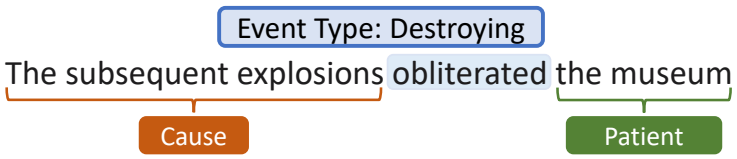

Figure 3.2: An illustration of EAE for the Destroying event comprising argument roles of Cause and Patient.

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.

3.2.1 Task Definition

In this work, we follow an alternate definition of **event** as a class attribute with values such as *occurrence, state, or reporting* [172, 67]. Following the early works of MUC [204, 61], **event arguments** are defined as participants in the event that provide specific and salient information about the event. **Event argument role** is the semantic category of the information the event argument provides. We illustrate in Figure 3.2 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. [3] For example, in Figure 3.2, *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] www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-entities-guidelines-v6.6.pdf

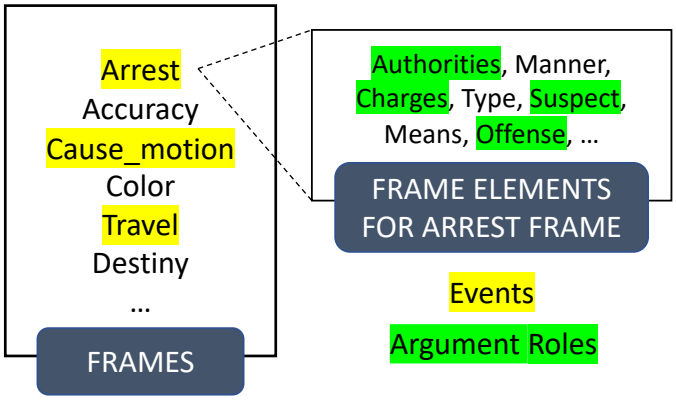

Figure 3.3: Illustration of challenges in using FrameNet for EAE - Not all frames are events, and not all frame elements are argument roles.

3.2.2 FrameNet for EAE

To overcome the challenge of constructing an event ontology from scratch, we leverage FrameNet, a semantic role labeling (SRL) dataset, to help our ontology creation. The similarity between SRL and EAE [3] provides us with the ground for leveraging FrameNet. SRL assigns semantic roles to phrases in the sentence, while EAE extracts event-specific arguments and roles from the sentence. Hence, *selecting event-related parts* of a fine-grained SRL dataset can be considered as an exhaustively annotated resource for EAE.

We choose FrameNet[4] [14] as the auxiliary SRL dataset since it is one of the most comprehensive SRL resources. It comprises 1200+ semantic frames [52], where a **frame** is a holistic background that unites similar words. [5] 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 § 3.2.1. We highlight these shortcomings in Figure 3.3, which enlists some FrameNet frames and frame elements for the *Arrest* frame. Based on EAE

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

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

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.

3.2.3 Building the EAE Ontology

To overcome the shortcomings of the basic mapping, we follow a two-step approach (Figure 3.4) and describe them in detail below.

Event Ontology To build the event ontology, we utilize the event mapping designed by MAVEN [220], 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 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. Annotators are provided with a list of frame elements along with their descriptions for each frame in the event ontology. [6] They are also provided with definitions for events and argument roles as discussed in Section 3.2.1. Based on these definitions, they are asked to annotate each frame element as (a) not an argument role, (b) an 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 super-generic frame elements (e.g., Time, Place, Purpose) unless they are relevant to the event. 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.

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

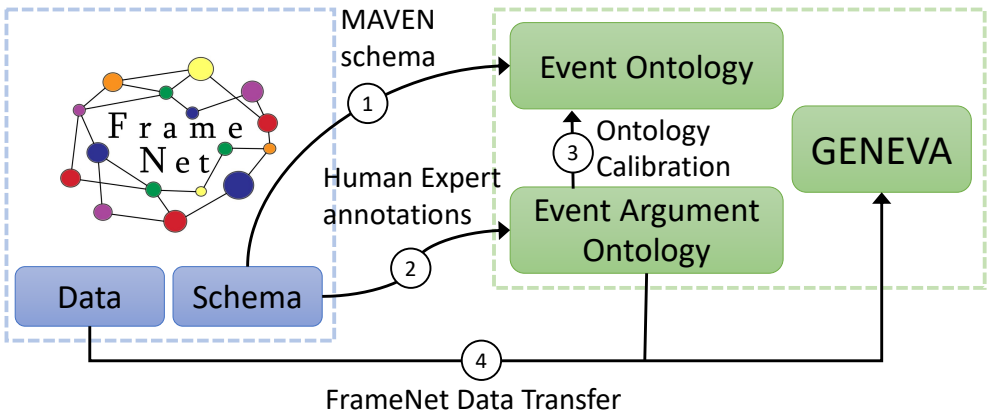

Figure 3.4: Illustration of the GENEVA creation from FrameNet labeled sequentially by the crucial steps. First, an event ontology is created using the MAVEN schema. Then, human expert annotations are used to create the event argument ontology. The event ontology is then calibrated using the argument ontology. Finally, data is transferred from FrameNet to GENEVA.

We recruited 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 [137] 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.

	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 3.1: 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.

3.2.4 Ontology Statistics

We present the statistics of our full ontology in Table 3.1 and compare it with existing ACE [37] and RAMS [45] ontologies. We use a subset of this Full ontology for creating GENEVA and include the statistics of the GENEVA ontology in the last column in Table 3.1. 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 [220], we show that our ontology spans 5 different abstract types and is the most diverse. 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 more comprehensive range of argument roles than other ontologies following ACE definitions of entity-centric argument roles.

In Figure 3.5, we organize our ontology into a hierarchy of the five abstract event types - Action, Change, Scenario, Sentiment, and Possession - as mentioned in the hierarchical tree structure introduced in MAVEN[220]. We show the extent of the overlap of the mapped ACE events in the GENEVA event schema (text labels colored in red). [7] We can observe that although there is some overlap between the datasets, GENEVA brings in a vast pool of new event types.

[7]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.

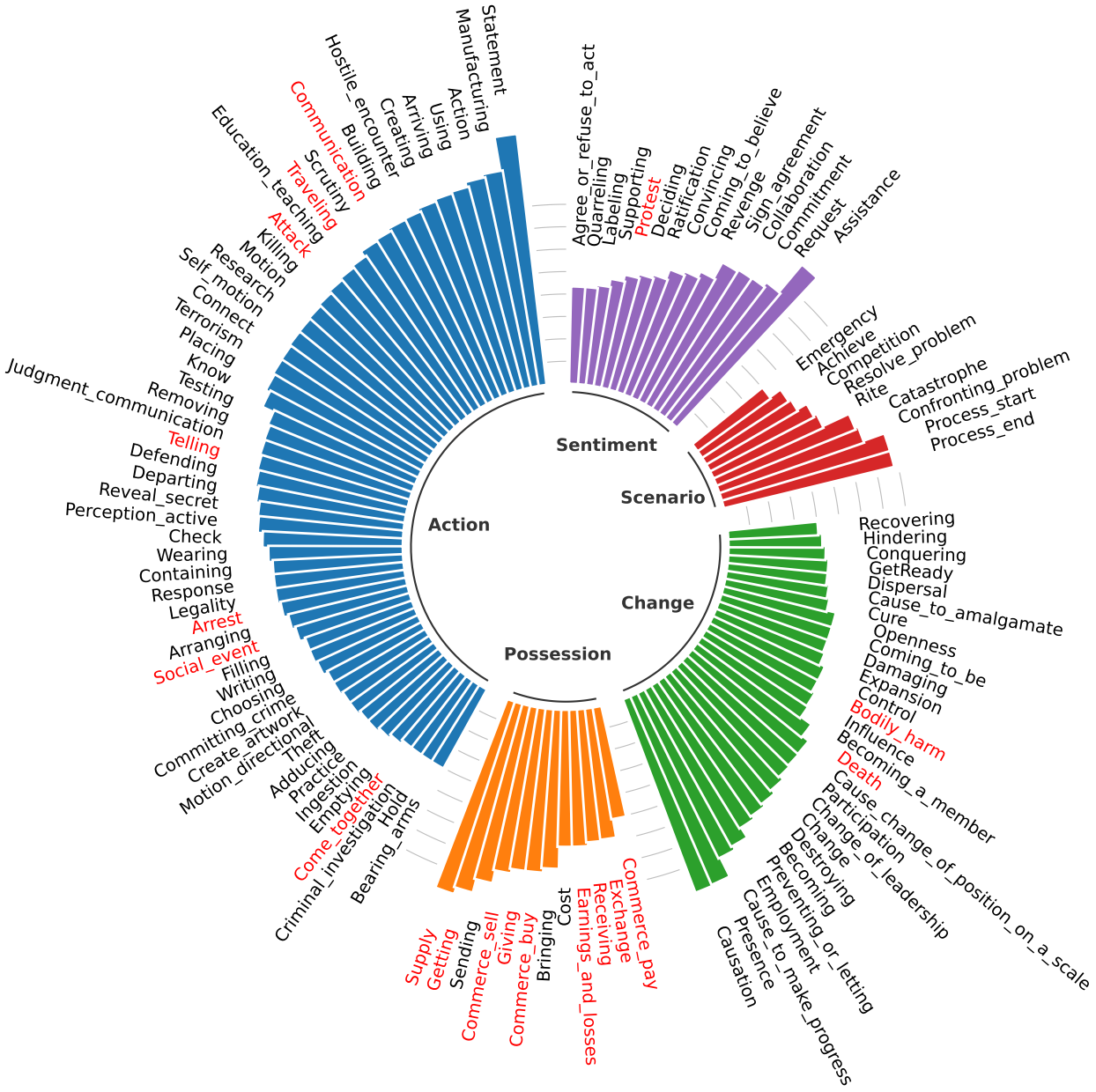

Figure 3.5: Circular bar plot for the various event types present in the GENEVA dataset organized into abstract event types. The height of each bar is proportional to the number of event mentions for that event (height is in log scale). Bar labels colored in red are the set of overlapping event types mapped from the ACE dataset.

3.3 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.

3.3.1 Dataset Creation

Since annotating EAE data for our large ontology is an expensive process, we leverage the annotated dataset of FrameNet to create GENEVA (Figure 3.4). 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. 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, as shown in Table 3.1.

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. 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 [54].

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. This assessment was done by two experts over 100 sampled annotations with 0.93 IAA measured by Cohen's kappa [137].

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 3.2: 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.

3.3.2 Data Analysis

The major statistics for GENEVA are shown in Table 3.2 along with its comparison with ACE and ERE.

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 3.1 further depicts how ACE/ERE focuses 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 3.2) is relatively less for GENEVA. Consequently, EAE models need to train from fewer examples on average which makes training more challenging.

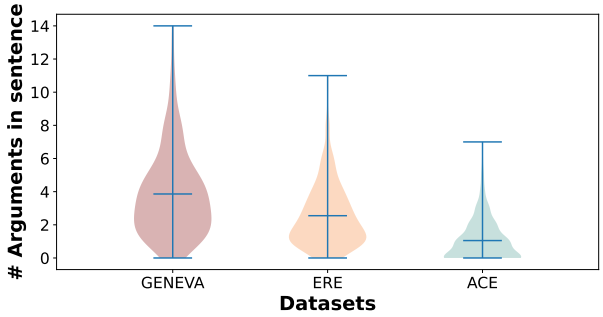

Figure 3.6: Violin plots for number of arguments per sentence for ACE, ERE, and GENEVA datasets.

Dense We plot the distribution of arguments per sentence [8] for ACE, ERE, and GENEVA in Figure 3.6. 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 From Figure 3.5, 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.

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.

3.3.3 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. 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 ($n = 5$).

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

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.[9] 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.

3.4 Benchmarking Results

3.4.1 Models and Metrics

Overall, we benchmark six EAE models from various representative families:

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++** [216], (2) **OneIE** [125], (3) **Query&Extract** [218].

Question-Answering models: Several works formulate event extraction as a machine reading comprehension task. We consider two models: (4) **BERT_QA** [42], (5) **TE** [133].

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** [159], (7) **DEGREE** [73].

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

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

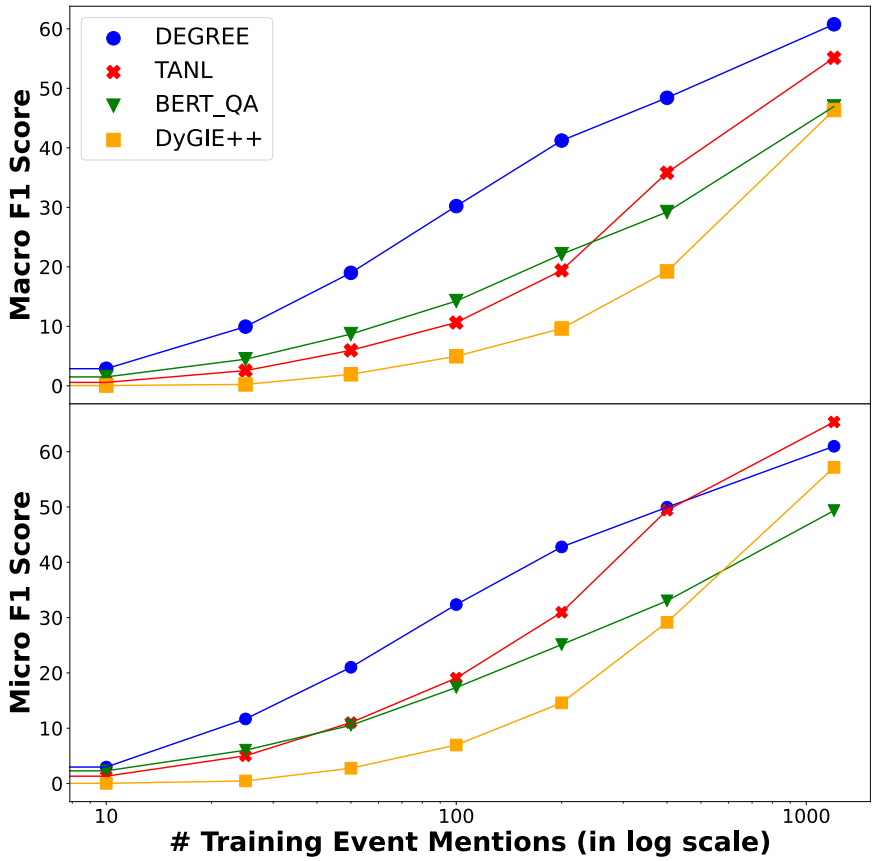

Figure 3.7: Model performance in macro F1 (top) and micro F1 (bottom) scores against the number of training event mentions (log-scale) for the low resource suite. Each datapoint is an average of 5 runs.

limited data test suites, we record a model performance curve, wherein we plot the F1 scores against the number of training instances.

3.4.2 Limited Training Data

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. Due to their inferior performance, we do not include OneIE and Query&Extract in the benchmarking results.

We present the model benchmarking results in terms of macro and micro F1 scores for the

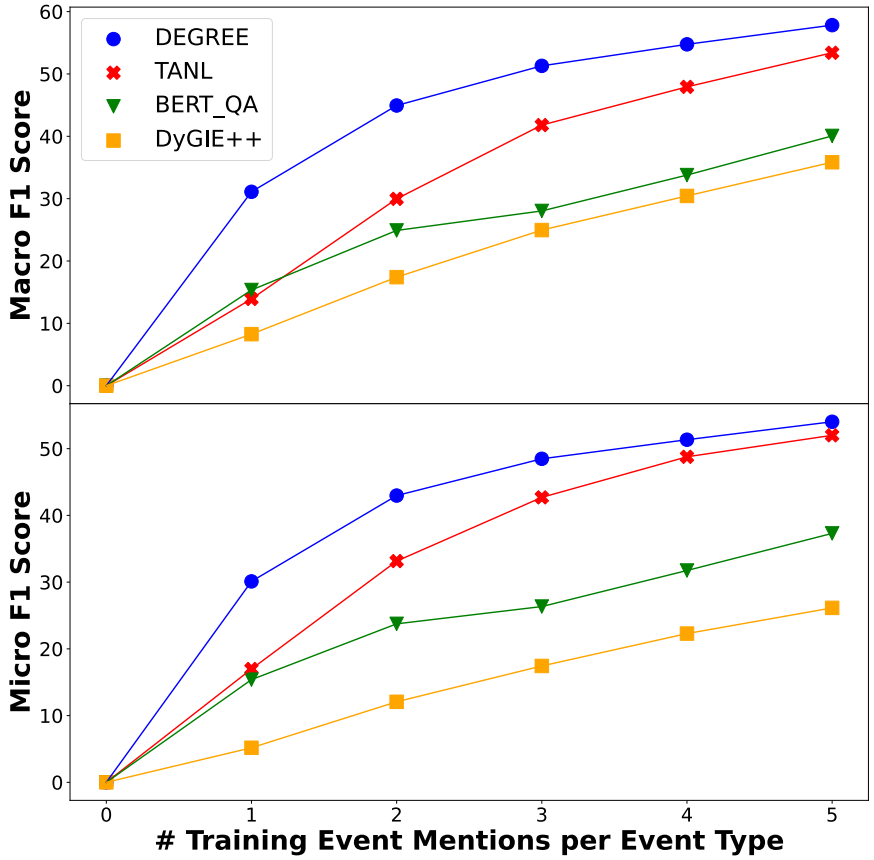

Figure 3.8: Model performance in macro F1 (top) and micro F1 (bottom) scores against the number of training event mentions per event for the few-shot suite. Each datapoint is an average of 5 runs.

low resource test suite in Figure 3.7 and for the few-shot test suite in Figure 3.8 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 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 biased toward high data events and do not generalize well uniformly across all events.

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 3.3: 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.

3.4.3 Unseen Event Data

We collate the results in terms of micro F1 scores for both the test suites in Table 3.3. 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 3.3, 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 is a pure zero-shot model (without training on any data), its performance is relatively super low in both settings. Thus, DEGREE shows superior transferability to unseen event types and argument roles.

3.4.4 Analysis

In this section, we provide analyses and case studies highlighting the insights and new challenges introduced by GENEVA.

Large Language Model Performance Recently, there has been an advent of Generative AI in the form of Large Language Models (LLMs) like GPT-3/4 [18], PaLM [26], Code4Struct [221], 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. More specifically, we provide 5 in-context examples from top-10 events and evaluate test examples from the remaining 105 events. Our GPT-prompt

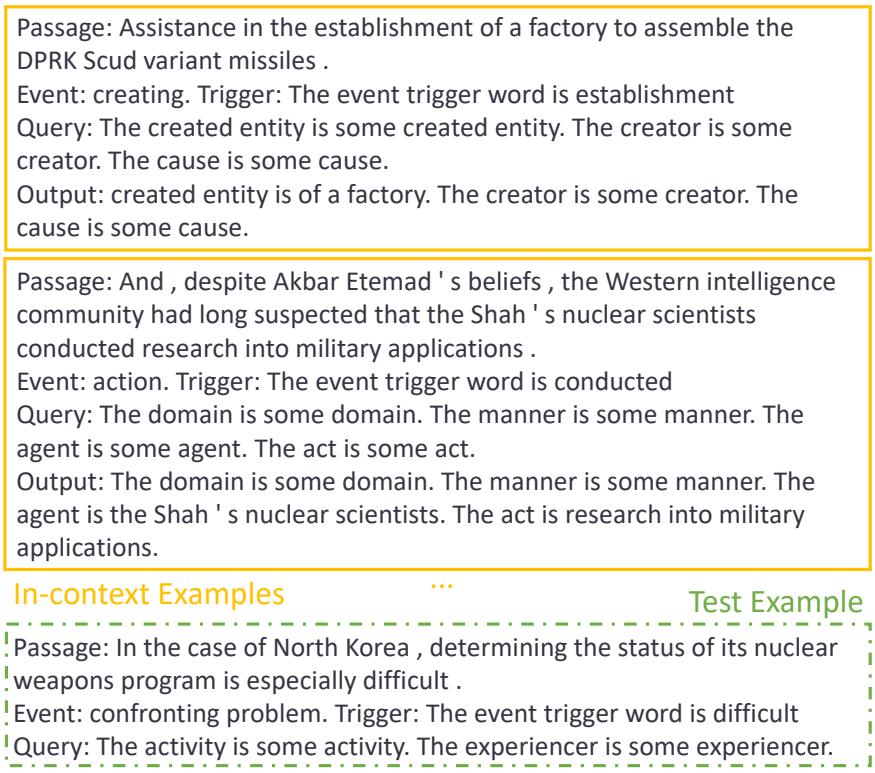

Figure 3.9: Illustration of the prompt used for evaluating GPT3.5-turbo. We provide 5 in-context examples (solid yellow) and provide the test example (dashed green).

template follows the DEGREE template wherein model replaces placeholders with arguments if present, else copies the original template. An illustration is provided in Figure 3.9.

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 are not directly comparable, it shows how GENEVA is quite challenging for LLMs in the zero-shot/few-shot setting.

New Challenge of Non-entity Roles In Table 3.4, 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 3.5. This ablation reveals a stark drop of 10-14% F1 points across all models when predicting non-entity

	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 3.4: 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. [73].

	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 3.5: 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.

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.

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 and provide results of the models on the full dataset in Table 3.6. Compared to the 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.

Case Study: Is ACE diverse enough? We conduct a case study to analyze how the limited diversity of ACE [37] can affect the generalizability of EAE models. We compare the

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

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 3.7: 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.

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 3.7.

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 ACE pre-trained 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.

3.5 Conclusion

Overall, in our work for this chapter, 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.

Chapter 4

CLAP: Contextual Label Projection for Cross-Lingual Structured Prediction

One important aspect to further generalizability is to extend information extraction models to a wider set of languages and, in turn, make these technologies accessible to a wider population. Cross-lingual transfer for structured prediction tasks such as named entity recognition, relation extraction, and event extraction has gained considerable attention recently [6, 85, 207, 53, 88, 21, 90]. In general, cross-lingual transfer generalizes models trained in a source language (typically English) to other target languages, broadening the scope of these applications to more languages [201, 23, 171].

One effective and simple way to improve cross-lingual transfer performance is translate-train, which leverages machine translation to generate pseudo-training data in the target languages by translating source language training data [233, 180, 246]. However, applying this technique to structured prediction tasks necessitates a *label projection* step, which involves jointly translating input sentences and labels [24]. Label projection requires not only *accurate translation* of the labels but also *maintaining the association* between the translated texts and labels. As illustrated in Figure 4.1, while "*suits*" can have multiple valid translations, only "诉讼" is presented in the translated sentence and is a proper translation at the same

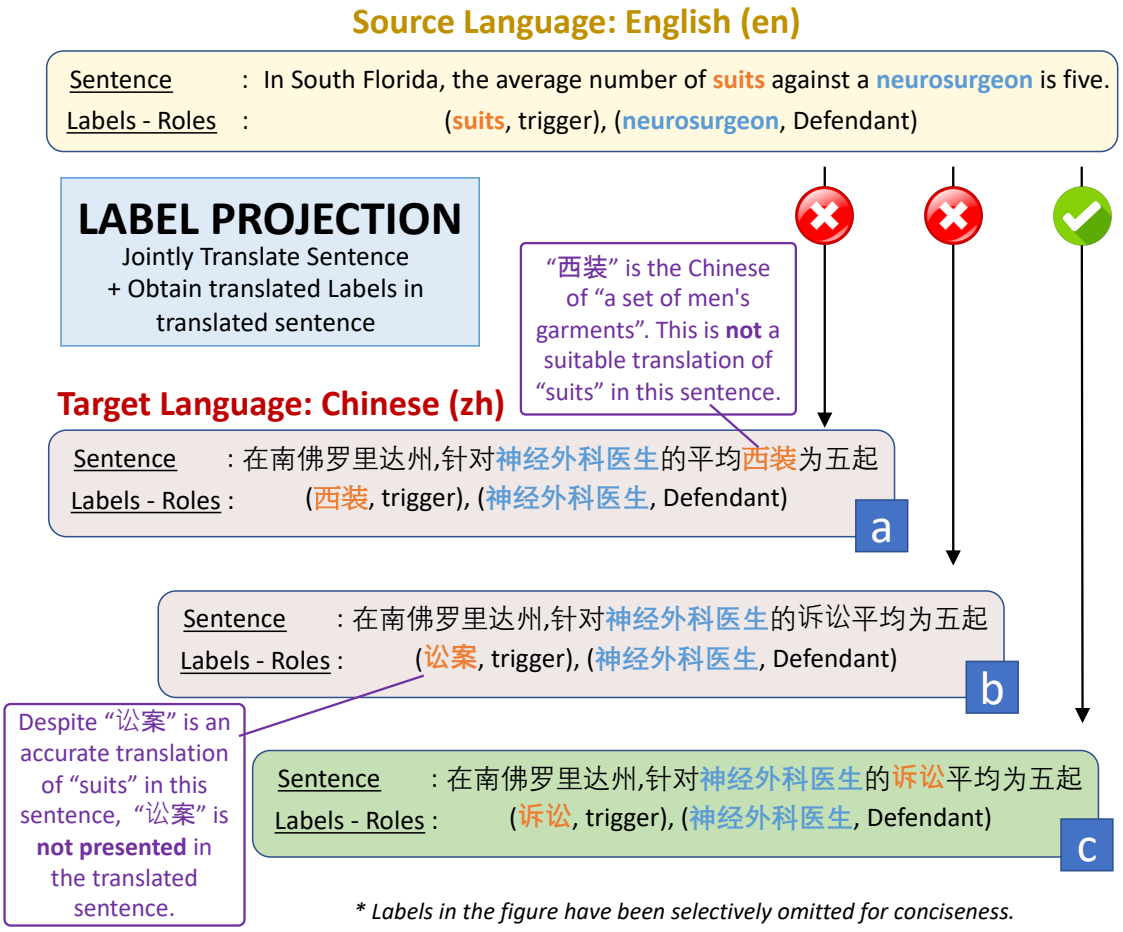

Figure 4.1: Illustration of the task of *label projection* from English to Chinese. Label projection converts sentences from a source to a target language while translating the associated labels jointly. Failures in this process occur when labels are either inaccurately translated or missing in the translated sentence in the target language.

time. This rigorous requirement on translating labels is different from translate-train for basic classification tasks, where mere sentence translation suffices [79].

Prior works have dealt with label projection through two primary frameworks. The first one, illustrated in Figure 4.2(a), performs machine translation on modified source sentences that incorporate label annotations using special markers [24, 70]. Translated labels can be extracted if special markers are retained in the translations; thus is contingent on the machine translation model. In this approach, the quality of the translation is *inherently compromised* due to the inclusion of special markers [24]. The other framework uses word similarity to procure word alignments between the source and translated sentences. Label translations

are further constructed by combining mapped tokens in the translated sentence [8, 200, 11], as shown in Figure 4.2(b). However, it is hard for this framework to ensure *accurate* label translation by merely using word alignments.

In this chapter, we introduce our work on CLAP (**C**ontextual **La**bel **P**rojection), which obtains projected label annotations by utilizing contextual machine translation for the labels. We first acquire the translation of the whole input sentence by any plug-and-play machine translation model. Then, inspired by the idea of contextual machine translation [215, 228], we use the translated input text as context to perform label translation, as shown in Figure 4.2(c). Exploiting contextual machine translation strongly enhances the *accuracy* of the translated labels while preserving their *association* to the translated sentence. Furthermore, translating the input sentence in an unmodified manner better exploits machine translators, and in turn, assures high quality of the translated sentence.

To implement contextual machine translation, we utilize an instruction-tuned language model with multilingual capabilities, Llama-2 [214]. We encode the translated input sentence and the constraint for the presence of labels in the form of instruction prompts. Despite sacrificing some translation ability compared to supervised machine translation models [255], instruction-tuned language models provide better understanding of contextual constraints.

We experiment on the tasks of event argument extraction (EAE) and named entity recognition (NER) using the ACE dataset [37] and the WikiANN dataset [158], covering 39 different languages in total. Our experiments show that utilizing label-projected data from CLAP for translate-train yields an average improvement of 1.7 and 1.4 F1 scores over strong baselines for EAE and NER, respectively. We also perform an intrinsic evaluation using a human study in Chinese, Arabic, Hindi, and Spanish to assess the projected labels' quality, which shows how CLAP provides more accurate label translations while preserving the label presence in the translated sentence. Further analyses also reveal how CLAP generalizes for different translation models and works effectively for the translate-test paradigm as well. These evaluations and robust analyses underscore the effectiveness of CLAP for label

projection.

We start by discussing the related works and providing a strong technical background about zero-shot cross-lingual structured prediction in Section 4.1. Next, we describe the major families of related works technically and introduce our work CLAP in Section 4.2. Then we discuss our experimental setup, baselines, and show the results in the form of intrinsic and extrinsic evaluations in Section 4.3. Finally, we provide an extensive set of analyses to demonstrate that our CLAP generalizes and has potential to improve further in Section 4.4. The work in the chapter has been published in Parekh et al. [162].

4.1 Background and Related Work

Here, we first discuss related works in this direction and later provide a technical description of the task of structure prediction.

Zero-shot Cross-lingual Structure Extraction Since the emergence of strong multilingual models [34, 29], various works have focused on zero-shot cross-lingual learning [79, 180] for various structure extraction tasks like named entity recognition [111, 237], relation extraction [148, 201], slot filling [101], and semantic parsing [150, 192]. Recent works have focused on building datasets [171, 161] as well as developing novel modeling designs exploring the usage of parse trees [201, 4, 76], data projection [242], pooling strategies [2], and generative models [73, 85] to improve cross-lingual transfer. We utilize the state-of-the-art model X-Gear [85] and XLM-R [29] as the downstream models for EAE and NER, respectively, and improve them further using CLAP-guided translate-train.

Label Projection Techniques Several works have attempted to solve label projection for various structure extraction tasks such as semantic role labeling [10, 50], slot filling [232], semantic parsing [143, 12], NER [149, 200], and question-answering [107, 110, 17]. The earliest works [243, 8] utilized statistical word-alignment techniques like GIZA++ [151] or

fast-align [44] for locating the labels in the translated sentence. Recent works [24] have also explored the usage of neural word aligners like QA-align [145] and Awesome-align [39]. Another set of works has explored the paradigm of mark-then-translate using special markers like quote characters (“") [110], XML tags (<a>) [79], and square braces ([0]) [24] to locate the translated labels. Overall, both these techniques can be error-prone and have poorer translation quality [8], as we show later in this chapter.

4.1.1 Technical Background

Structure Prediction Task

Given an input sentence $\mathbf{x}$, structure prediction models aim to predict structure output $\mathbf{y} = [\mathbf{x}[i_1 : j_1], \mathbf{x}[i_2 : j_2], \ldots, \mathbf{x}[i_n : j_n]]$ (where $\mathbf{x}[i_1 : j_1]$ is an input sentence span from token i_1 to j_1) corresponding to a set of roles $\mathbf{r} = [r_1, r_2, \ldots, r_n]$ (where $r_i \in \mathcal{R}$, a pre-defined set of roles). This vastly differs from standard classification-based tasks wherein the output prediction y is a singular value from a fixed set of classes independent of the input $\mathbf{x}$.

Zero-shot Cross-Lingual Transfer

Zero-shot cross-lingual transfer [5, 79, 84, 75] aims to train a downstream model for the target language l_{tgt} using supervised data $\mathcal{D}_{src}$ from a source language l_{src} without using any data in the target language (i.e. $\mathcal{D}_{tgt} = \phi$). The paradigm has effectively advanced language technologies for under-resourced languages, as shown in Hu et al. [79] and Ruder et al. [180].

Translate-Train

Translate-train [79, 180] is a popular and powerful zero-shot cross-lingual transfer technique that leverages machine translators $\mathcal{T}$ to boost downstream model performance. Specifically, in translate-train, $\mathcal{D}_{src}$ is translated into the target language as pseudo training data $\mathcal{D}_{src}^{tgt}$ and the downstream model is trained using a combination of $\{\mathcal{D}_{src}, \mathcal{D}_{src}^{tgt}\}$.

Utilizing translate-train for structured prediction tasks requires *Label Projection*, which includes two sets of translations: (1) Sentence translation ($\mathbf{x}^{src} \xrightarrow{\mathcal{T}} \mathbf{x}^{tgt}$), where we use $\xrightarrow{\mathcal{T}}$ to denote the translation from l_{src} to l_{tgt} using $\mathcal{T}$; and (2) Label translation ($\mathbf{y}^{src} \rightarrow \mathbf{y}^{tgt}$), such that the translated label $\mathbf{y}^{tgt}$ is appropriately *associated with* $\boldsymbol{x}^{tgt}$. This demand makes translate-train for structure prediction tasks more complex than that for classification tasks, as the latter only requires sentence translation (since y is a value independent of $\mathbf{x}$). [1]

Translate-Test

Besides translate-train, translate-test is another commonly used technique in zero-shot cross-lingual transfer. During testing time, it uses models solely trained on $\mathcal{D}_{src}$ to make predictions on translated test sentences ($\mathbf{x}^{tgt} \xrightarrow{\mathcal{T}} \mathbf{x}^{src}$), and then uses label projection to map predictions on $\mathbf{x}^{src}$ back to predictions on $\mathbf{x}^{tgt}$. Since it will cause additional error propagation issues during inference time, we mainly focus on translate-train in this paper. However, we discuss CLAP's utilization and effectiveness on translate-test as a case study in § 4.4.5.

Label Projection

We hereby technically define the problem of *label projection* [8, 24]:

$$\begin{aligned} & \mathbf{x}^{src} \xrightarrow{\mathcal{T}} \mathbf{x}^{tgt} & \\ \& \quad & y_m^{src} \rightarrow y_m^{tgt} & \forall y_m^{src} \in \mathbf{y} \\ s.t. \quad & y_m^{tgt} \in \mathbf{x}^{tgt} & \forall y_m^{tgt} \in \mathbf{y}^{tgt}. \end{aligned}$$

This problem requires optimizing two properties of **accuracy** and **faithfulness** on the translations. Accuracy ensures that $[\mathbf{x}^{tgt}, y_1^{tgt}, y_2^{tgt}, \ldots, y_n^{tgt}]$ are accurate translations of $[\mathbf{x}^{src}, y_1^{src}, y_2^{src}, \ldots, y_n^{src}]$. On the other hand, faithfulness ensures that each y_m^{tgt} is associ-

[1] For certain structure prediction tasks like relation classification (determining the relationship between two entities in $\mathbf{x}$), even if the output y is scalar, translate-train necessitates a label projection step due to the required projection of the two given entities into the translated sentence.

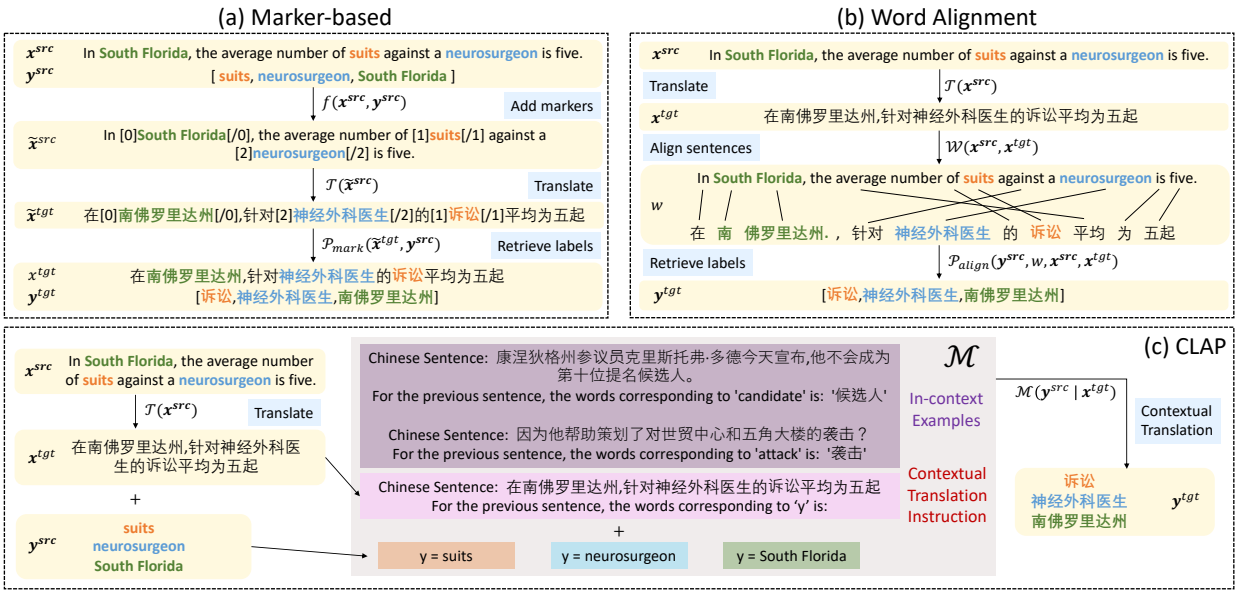

Figure 4.2: Illustration of the various techniques to conduct label projection: (a) **Marker-based Translation** use markers to transform the sentence and translate the transformed sentence with label markers jointly, (b) **Word Alignment** methods use external word alignment tools to locate the translated labels in the translated sentence, and (c) **CLAP** performs contextual translation on labels using $\mathcal{M}$ (here we show instruction-tuned language model as $\mathcal{M}$) to locate translated label in the translated sentence.

ated with $\mathbf{x}^{tgt}$ (the constraint of $y_m^{tgt} \in \mathbf{x}^{tgt}$). Standard translation models $\mathcal{T}$ trained on supervised sentence translation pairs cannot simply impose the additional faithfulness constraint, such as the failure cases shown in Figure 4.1. This demonstrates the challenges of the label projection.

4.2 Methodology

Here, we first formally define the previous attempts at label projection and later introduce CLAP, which provides a new perspective on using contextual machine translation for label projection.

4.2.1 Baseline Methods

As stated above, two primary frameworks, namely, marker-based translations and word-alignment-based methods, are primarily used in prior works.

Marker-based Translations solve the label projection by first marking labels to the input sentence $\mathbf{x}^{src}$, forming $\tilde{\mathbf{x}}^{src}$, and then use the translation model to obtain the potential translation of the input sentence and labels jointly [110, 79, 24]. For example, in Figure 4.2(a), "South Florida" is delineated by markers [0] and [\0]. Assuming the preservation of markers after translation of $\tilde{\mathbf{x}}^{src}$, a post-processing step, $\mathcal{P}_{mark}$, is performed to retain the translated labels $\mathbf{y}^{tgt}$ and translated sentence $\mathbf{x}^{tgt}$. Putting every step together, we have

$$\tilde{\mathbf{x}}^{src} = f(\mathbf{x}^{src}, \mathbf{y}^{src}), \quad \tilde{\mathbf{x}}^{tgt} = \mathcal{T}(\tilde{\mathbf{x}}^{src})$$
$$\mathbf{x}^{tgt}, \mathbf{y}^{tgt} = \mathcal{P}_{mark}(\tilde{\mathbf{x}}^{tgt}, \mathbf{y}^{src}),$$

where f denotes the marker addition step and $\tilde{\mathbf{x}}^{tgt}$ is the translation of $\tilde{\mathbf{x}}^{src}$ using translator $\mathcal{T}$.

Despite their simplicity, these methods suffer from poor translation quality and reduced robustness to different translation models owing to their input sentence transformations and strong assumptions about the retention of markers in $\tilde{\mathbf{x}}^{tgt}$.

Word Alignment approaches [8, 242] first translate the input sentence and acquire word alignments [44, 39] between the translation pairs. Each translated label y_m^{tgt} is then procured by merging the aligned words of y_m^{src} in the translated sentence using the word mappings w. For example, in Figure 4.2(b), the translated label for "South Florida" is obtained by merging two aligned words, which is done by a heuristic post-processing algorithm $\mathcal{P}_{align}$. Formally,

we have

$$\mathbf{x}^{tgt} = \mathcal{T}(\mathbf{x}^{src}),\ w = \mathcal{W}(\mathbf{x}^{src}, \mathbf{x}^{tgt})$$
$$y_m^{tgt} = \mathcal{P}_{align}(y_m^{src}, w, \mathbf{x}^{src}, \mathbf{x}^{tgt}) \qquad \forall y_m^{src} \in \mathbf{y}^{src}$$

Although these approaches provide high-quality sentence translations, their translated labels can be error-prone as they use simple word alignment modules for capturing word-level translation relations without considering the entire label for translation [8, 24].

4.2.2 CLAP

We tackle the task of label projection through a new perspective - performing actual translation on labels instead of recovering them from translated text $\mathbf{x}^{tgt}$. This better ensures the accuracy of the translated labels $\mathbf{y}^{tgt}$. To accomplish this, we leverage the idea of *contextual machine translation* on the label translation with $\mathbf{x}^{tgt}$ as context.

Contextual machine translation, which aims to perform phrase-level translations conditional on the context of the translated sentence, is tangentially explored for applications like anaphora resolution [215] and pronoun translations [228]. The main goal of this task is to maintain the consistency of phrasal translations in the given context. In our work, we develop a novel model, CLAP, to extend the idea of contextual translation to the application of label projection.

As illustrated in Figure 4.2(c), CLAP first utilizes machine translation model $\mathcal{T}$ to translate input sentence $\mathbf{x}^{src}$ to $\mathbf{x}^{tgt}$. Treating $\mathbf{x}^{tgt}$ as the context, the contextual translation model $\mathcal{M}$ translates the labels $\mathbf{y}^{src}$ to $\mathbf{y}^{tgt}$. Contextual translation implicitly imposes the *faithfulness constraint* which requires $y_m^{tgt} \in \mathbf{x}^{tgt} \quad \forall y_m^{tgt} \in \mathbf{y}^{tgt}$, hence, slackly satisfying the

requirement of label projection. These two steps can be formally described as

$$\mathbf{x}^{tgt} = \mathcal{T}(\mathbf{x}^{src})$$
$$y_m^{tgt} = \mathcal{M}(y_m^{src}|\mathbf{x}^{tgt}) \qquad \forall y_m^{src} \in \mathbf{y}^{src}$$

where y_m^{tgt} is generated from $\mathcal{M}(y_m^{src}|\mathbf{x}^{tgt})$, drawing the significant difference from the previous works.

Compared to word alignment approaches using simple word-similarity aligners $\mathcal{W}$, we use models with translation capabilities $\mathcal{M}$ to improve the accuracy of translated labels. Furthermore, the independence of $\mathcal{T}$ and $\mathcal{M}$ for translating $\mathbf{x}^{src}$ and $\mathbf{y}^{src}$ respectively, assures that CLAP has better translation quality for $\mathbf{x}^{tgt}$ and is more robust than the marker-based baselines. We empirically back these intuitions in § 4.3.3.

4.2.3 Implementing CLAP

Putting our idea into practice, we configure $\mathcal{T}$ to be a modular component that can be replaced by any third-party translation model. For $\mathcal{M}$, we use an instruction-tuned language model with multilingual capabilities. Instruction-tuned language models can accept conditional information in their natural language prompts. Specifically, we encode the translated target sentence $\mathbf{x}^{tgt}$ as well as the faithfulness constraint $y_m^{tgt} \in \mathbf{x}^{tgt}$ implicitly in the form of natural language instructions (highlighted as "Contextual Translation Instruction" in Figure 4.2(c)). Following Brown et al. [18], we also provide n randomly chosen in-context examples (highlighted as "In-context examples" in Figure 4.2(c)) to improve the instruction-understanding capability of the model. [2] Lastly, we use simple string-matching algorithms to get the exact span index of y_m^{tgt} in $\mathbf{x}^{tgt}$. Although this may not be the optimal solution when duplicated strings exist in $\mathbf{x}^{tgt}$, it works well in practice as stated in prior word-alignment methods [39].

[2]The in-context examples are generated using Google translation and initial prediction from instruction-tuned LMs. The label predictions are further verified by back-translation.

4.3 Experiments and Results

This section describes our experimental setup comprising the datasets, baselines, and implementation details. Later, we present both intrinsic and extrinsic evaluations of CLAP.

4.3.1 Task and Dataset

We choose two structure prediction tasks, event argument extraction (EAE) [204, 74] and named entity recognition (NER) [210, 211] for evaluating our label projection method. EAE requires the extraction of text segments serving as arguments corresponding to an event and mapping them to their corresponding argument roles. NER aims to identify and categorize named entities from the input sentence. We use the multilingual ACE dataset [37] and the WikiANN [158, 175] for benchmarking EAE and NER, respectively. We consider the zero-shot cross-lingual transfer using English (en) as the source language for both tasks. For ACE, we follow the pre-processing by Huang et al. [85] to retain 33 event types and 22 argument roles. For WikiANN, we utilize pre-processing by Hu et al. [79]. We provide the high-level statistics for these datasets in Table 4.1.

4.3.2 Baselines and Implementation Details

We select two label projection models as baselines, each representing the two baseline frameworks we covered in Section 4.2.1, respectively: (1) **EasyProject** [24], a recent marker-based translation technique, utilizes numbered square braces (e.g., [0] and [/0]) to mark the labels in the input sentence. (2) **Awesome-Align** [39], a neural bilingual word alignment model, uses multilingual language models to find word similarities to derive word alignments, which are later used for label projection.

For the translation model $\mathcal{T}$, we experiment with the Google Machine Translation (GMT).[3] For CLAP, we use the text-completion version of Llama-2 [214] with 13B parameters. as

[3] https://cloud.google.com/translate. We use the free scraping tool to reduce the translation costs.

	ACE	**WikiANN**
# Train Instances	4,202	20,000
# Dev Instances	450	10,000
# Avg. Test Instances	194	6,469
# Test Languages	2	39
Test Languages	Arabic (ar), Chinese (zh)	Arabic (ar), Bulgarian (bg), Bengali (bn), German (de), Greek (el), Spanish (es), Estonian (et), Basque (eu), Farsi (fa), Finnish (fi), French (fr), Hebrew (he), Hindi (hi), Hungarian (hu), Indonesian (id), Italian (it), Japanese (ja), Javanese (jv), Georgian (ka), Kazakh (kk), Korean (ko), Malayalam (ml), Marathi (mr), Malay (ms), Burmese (my), Dutch (nl), Portuguese (pt), Russian (ru), Swahili (sw), Tamil (ta), Telugu (te), Thai (th), Tagalog (tl), Turkish (tr), Urdu (ur), Vietnamese (vi), Yoruba (yo), Chinese (zh)

Table 4.1: High-level data statistics for ACE and WikiANN datasets for EAE and NER tasks, respectively. # = 'number of' and Avg. = 'average'.

$\mathcal{M}$. We use $n = 2$ in-context examples for CLAP prompts. For Awesome-align, we use the unsupervised version of their model utilizing multilingual BERT [34] as it provides better results [24].

4.3.3 Intrinsic Evaluation

We first evaluate CLAP by directly evaluating the label projection quality, mainly focusing on evaluating the accuracy and faithfulness of the translated labels, with the definition stated in Section 4.1.1.

Accuracy is measured in terms of translation quality by comparing the source labels with the translated labels ($\mathbf{y}^{src} \leftrightarrow \mathbf{y}^{tgt}$). We hire 5 native bilingual speakers to rank the translated labels by the different models based on their translation quality. The native speakers were undergraduate and graduate students who were well-versed in their respective native languages. We conduct this evaluation on 50 data samples for four languages - Chinese, Arabic, Hindi, and Spanish, respectively. The final accuracy score for each model is the average percentage when the given methods provided the best quality translation for the

System 1	v/s	System 2	Arabic			Chinese		
			S1	**Tie**	**S2**	**S1**	**Tie**	**S2**
CLAP		Awesome-align	**36%**	58%	6%	**45%**	50%	5%
CLAP		EasyProject	**52%**	32%	16%	**56%**	39%	5%
CLAP		Independent	18%	60%	**22%**	12%	71%	**17%**
Independent		Awesome-align	**44%**	42%	14%	**39%**	57%	4%
Independent		EasyProject	**50%**	44%	6%	**50%**	46%	4%
Awesome-align		EasyProject	**42%**	26%	32%	**34%**	50%	16%
			Hindi			**Spanish**		
			S1	**Tie**	**S2**	**S1**	**Tie**	**S2**
CLAP		Awesome-align	**20%**	74%	6%	**12%**	84%	4%
CLAP		EasyProject	**42%**	48%	10%	**30%**	66%	4%
CLAP		Independent	**18%**	64%	**18%**	**24%**	68%	8%
Independent		Awesome-align	**28%**	60%	12%	**20%**	64%	16%
Independent		EasyProject	**52%**	36%	12%	**32%**	52%	16%
Awesome-align		EasyProject	**42%**	42%	16%	**26%**	64%	10%

Table 4.2: A/B comparison of the various label projection techniques for accuracy evaluation for the Google Translation model. Accuracy is measured as the label translation quality by native human speakers. Here, **S1** = System 1 is better, **S2** = System 2 is better, and **Tie** = similar quality. The better systems are highlighted in **bold**.

labels among the other competitor models.

Faithfulness measures the fulfillment of the label projection constraint. It is measured as a percentage of projected data points when all the translated labels are present in the translated input sentence $\left(y_m^{tgt} \in \mathbf{x}^{tgt} \quad , \forall y_m^{tgt} \in \mathbf{y}^{tgt}\right)$. The statistics use the complete test set on ACE and WikiANN.

Results

We present the complete results for accuracy evaluation as an A/B comparison of the different techniques in terms of their win rates (i.e., percentage when A is better than B) in Table 4.2. We clearly observe how CLAP is more accurate than previous baselines of Awesome-align and EasyProject, while at par with the Independent baseline.

We present the complete results for the faithfulness evaluation per language in Tables 4.3

Techniques	ar	zh	Avg.
Independent	33	38	35
Awesome-align	66	83	74
EasyProject	31	66	48
CLAP	74	85	**79**

Table 4.3: Faithfulness evaluation of the various label projection techniques for EAE as a percentage of the times the translated labels were present in the translated input sentence. Numbers are in percentage (%). Higher faithfulness is better, and the best techniques are highlighted in **bold**.

and 4.4 for EAE and NER tasks, respectively. For EAE, CLAP has the best faithfulness, followed by Awesome-align. For NER, Awesome-align and EasyProject have the highest faithfulness.

The combined results for accuracy and faithfulness of the models are plotted together in Figure 4.3. An ideal model should optimize both these metrics and thus, the closer the models are to the top-right, the better they are deemed. Overall, this figure shows how CLAP performs the best intrinsically, as it is the closest to the top-right for both tasks. For EAE, CLAP is better than all models in both metrics, while for NER, CLAP compromises faithfulness slightly for stronger accuracy. Awesome-align and EasyProject are both great at attaining higher projection rates, but produce more inaccurate label translations. Overall, intrinsic evaluation reveals how CLAP provides the best balance of accuracy and faithfulness.

4.3.4 Extrinsic Evaluation

Extrinsic evaluation implicitly evaluates the label projection techniques' ability to generate good-quality data for downstream tasks. The projected data is utilized to train downstream models using the translate-train paradigm together with the original training data in English. For translate-train, we only retain the projected datapoints that satisfy the faithfulness constraint as part of the target pseudo-training data $\mathcal{D}_{src}^{tgt}$. Furthermore, we also compare these fine-tuned models with **LLM-Infer**, a large language model (LLM) baseline (LLama-2-

Techniques	af	ar	bg	bn	de	el	es	et	eu	fa	fi	fr	he	hi
Independent	78	66	67	74	79	57	70	70	64	61	71	71	71	65
Awesome-align	99	95	98	92	99	98	99	98	97	96	99	98	95	93
EasyProject	100	98	83	98	97	89	99	97	94	99	98	99	94	36
CLAP	94	75	63	93	79	46	84	92	91	72	92	74	80	90
	hu	**id**	**it**	**ja**	**jv**	**ka**	**kk**	**ko**	**ml**	**mr**	**ms**	**my**	**nl**	**pt**
Independent	68	77	74	68	66	64	56	63	57	73	80	53	76	76
Awesome-align	98	99	99	58	98	95	94	96	88	92	99	90	99	97
EasyProject	97	99	98	95	94	99	77	93	87	73	98	62	100	99
CLAP	93	84	78	67	53	70	85	64	88	95	82	55	85	89
	ru	**sw**	**ta**	**te**	**th**	**tl**	**tr**	**vi**	**ur**	**yo**	**zh**	**Avg.**		
Independent	59	79	72	76	66	81	76	74	74	45	66	69		
Awesome-align	97	96	91	91	51	99	98	83	97	92	92	**93**		
EasyProject	99	97	91	87	99	99	98	98	94	77	92	92		
CLAP	66	94	96	90	57	58	94	89	91	88	60	79		

Table 4.4: Faithfulness evaluation of the various label projection techniques for NER as a percentage of the times the translated labels were present in the translated input sentence. Numbers are in percentage (%). Higher faithfulness is better, and the best techniques are highlighted in **bold**.

13B-chat model) directly inferring in the target language.

EAE For the EAE downstream model, we use the state-of-the-art model for zero-shot cross-lingual EAE: X-Gear [85]. We explore three versions of the X-Gear model: mBART without copy (mBART), mT5 without copy (mT5), and mT5 with copy mechanism (mT5+Copy). We present the results in terms of argument classification F1 scores [4] in Table 4.5. For reference, we also include the zero-shot baseline (training only on $\mathcal{D}_{src}$). Evidently, CLAP performs the best, providing an average gain of 1.7 F1 points over the next best baseline of Awesome-align and a net gain of 7.5 F1 points over the zero-shot baseline. This result is in sync with our intrinsic evaluation, wherein CLAP performed the best for EAE.

NER For NER, we utilize XLM-RoBERTa$_{\text{large}}$ [29] as our downstream model and use the XTREME [79] setup for implementation. The main results for entity classification F1

[4]Averaged over five model runs

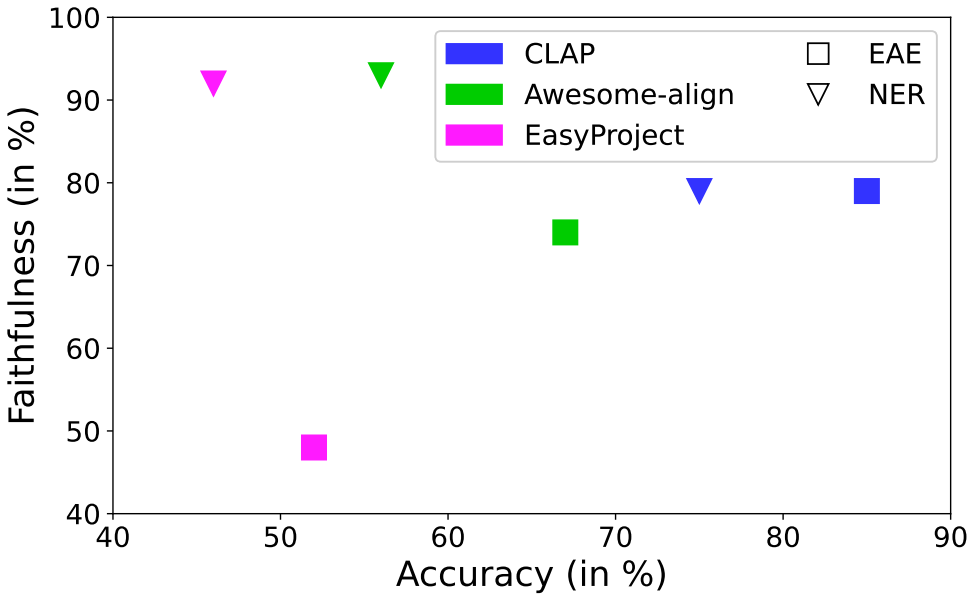

Figure 4.3: Reporting faithfulness and accuracy (in %) for the different label projection models on EAE and NER datasets. The closer the model is to the top-right, the better it is.

	mBART		**mT5**		**mT5+Copy**		**Avg**
	ar	zh	ar	zh	ar	zh	
Zero-shot*	36.3	47.3	36.7	51.0	40.3	51.9	43.9
LLM-Infer	-	-	-	-	16.9^{+}	24.0^{+}	20.5
Awesome-align	45.2	49.4	**46.8**	53.7	48.6	54.5	49.7
EasyProject	37.9	52.3	34.5	54.6	38.5	56.3	45.7
CLAP (ours)	**46.0**	**53.4**	44.3	**56.5**	**49.3**	**58.6**	**51.4**

Table 4.5: Extrinsic evaluation of the different label projection techniques regarding downstream model performance using translate-train for EAE. Avg = Average. * indicates the reproduced results of our base zero-shot cross-lingual EAE model, X-Gear [85]. Results for LLM-Infer (marked with $^{+}$) are independent of the XGear base model.

scores are presented in Table 4.6 along with the zero-shot baseline. Overall, CLAP performs the best with an absolute improvement of 0.7 F1 points over the zero-shot baseline and 1.4-1.7 F1 points over the previous works. The strong downstream model performance using CLAP combined with our learnings from intrinsic evaluation underscores the importance of prioritizing accuracy over faithfulness for NER.

Lang	af	ar	bg	bn	de	el	es	et	eu	fa	fi	fr	he	hi	hu
Zero-shot	77.4	48.1	82.8	77.0	78.8	80.6	74.5	78.7	61.4	69.2	79.3	79.4	57.3	70.6	80.8
LLM-Infer	50.9	24.8	66.9	12.0	44.2	42.2	59.5	41.6	36.7	19.5	46.7	53.5	15.6	18.9	20.6
Awesome-align	**77.9**	46.0	**81.0**	**81.2**	**78.8**	71.7	65.3	**78.0**	66.8	46.4	**77.4**	78.2	55.3	**73.9**	**77.4**
EasyProject	76.1	34.4	**81.0**	78.6	**78.8**	69.3	70.5	73.9	54.8	49.1	**77.8**	78.8	**61.1**	73.0	75.6
CLAP	74.4	**48.7**	**81.0**	78.1	78.4	**75.9**	**74.7**	77.4	**68.8**	**59.0**	75.9	**79.4**	58.4	73.1	72.4

	id	it	ja	jv	ka	kk	ko	ml	mr	ms	my	nl	pt	ru	sw
Zero-shot	53.1	79.4	19.1	58.5	72.3	51.9	57.5	66.4	65.3	53.4	65.8	83.0	80.0	74.2	68.4
LLM-Infer	30.3	56.0	35.7	28.7	21.7	20.9	18.5	11.1	16.5	46.5	10.1	64.3	46.4	22.7	33.4
Awesome-align	52.8	79.3	20.3	56.3	70.4	**47.7**	57.7	**63.4**	**62.4**	70.7	54.1	**83.0**	75.8	64.8	70.1
EasyProject	51.0	79.0	41.3	62.4	66.4	31.7	48.2	56.5	59.8	71.7	60.3	81.9	**79.6**	66.3	**71.5**
CLAP	**56.1**	**80.1**	**45.3**	**64.8**	**70.5**	42.8	**60.1**	60.3	61.4	**73.5**	**61.5**	82.2	78.2	**68.3**	70.6

	ta	te	th	tl	tr	ur	vi	yo	zh	Avg
Zero-shot	12.8	9.2	19.8	46.1	31.0	11.6	37.3	28.6	41.0	32.1
Awesome-align	**62.4**	**55.4**	2.4	**80.9**	62.8	53.7	66.4	**61.5**	45.4	63.5
EasyProject	53.2	54.2	11.4	78.2	**66.8**	**63.8**	65.6	68.8	42.0	63.2
CLAP	59.6	53.1	**13.2**	74.6	62.9	32.9	**75.8**	59.6	**49.7**	**64.9**

Table 4.6: Extrinsic evaluation of the different label projection techniques in terms of downstream model performance using translate-train for NER. Avg = Average.

4.4 Analysis and Case Studies

4.4.1 Qualitative Analysis

Diving deeper, we qualitatively study typical error cases for the translated labels in four languages by different label projection techniques. In 200 examples of our study, we found that 18% of the time, EasyProject predicts nothing due to markers dropped in the translated sentence, and for 19%, EasyProject simply copies the English label, failing to translate it to the target language. For Awesome-align, the majority of errors are due to additional words or incomplete label translations, similar to the observation presented in [24]. This could be because it is hard for the word-alignment module to decide alignments between sub-words, leading to over-alignment or under-alignment. We show two selected examples of our study from Hindi (hi) and Chinese (zh) in Table 4.7, where we show how Awesome-align predicts extra words or incomplete words owing to misalignments, and EasyProject fails to

Source Sentence	Source Label	Target Lang	Technique	Translated Label	Explanation
Born in Castelvetrano , Trapani and raised in Catania , he moved to Madrid to keep up his busy career .	Castelvetrano	hi	Awesome-align	कैस्टेलवेट्रानो ट्रापानी (Castelvetrano Trapani)	Extra word
			EasyProject	Castelvetrano	No translation
			CLAP	कैस्टेलवेट्रानो (Castelvetrano)	Perfect
Unilaterally leading a coalition featuring tyrannies, effect such change remains a bad idea, Iraq's elections notwithstanding.	Iraq	zh	Awesome-align	伊拉 (Ira-)	Incomplete
			EasyProject	尽管伊拉克 (although Iraq)	Extra word
			CLAP	伊拉克 (Iraq)	Perfect

Table 4.7: Qualitative examples highlighting the error cases of the baseline models along with explanations for Hindi (hi) and Chinese (zh). We also show how CLAP performs better and fixes the errors.

	mBART		mT5		mT5+Copy		Avg
	ar	zh	ar	zh	ar	zh	
Zero-shot	36.3	47.3	36.7	51.0	40.3	51.9	43.9
Awesome-align	**45.7**	48.6	43.1	52.1	47.1	53.8	48.4
EasyProject	37.3	**53.6**	35.3	54.0	36.5	55.6	45.4
CLAP (ours)	45.5	52.0	**44.8**	**54.7**	**48.2**	**56.9**	**50.4**

Table 4.8: Extrinsic evaluation of the different label projection techniques using translate-train for EAE using the mBART-50 many-to-many translation model.

translate the word for Hindi while producing extra tokens for Chinese. In both cases, we show how CLAP makes accurate predictions and is more robust in maintaining accurate label translations.

4.4.2 Generalization to other translation models

To verify the generalizability of our approach to other translation models, we perform an extrinsic evaluation of the label projection techniques on the EAE task using the mBART-50 many-to-many (MMT) [99] translation model. We show the results for this evaluation in

	mBART		mT5		mT5+Copy		Avg
	ar	zh	ar	zh	ar	zh	
Zero-shot	36.3	47.3	36.7	51.0	40.3	51.9	43.9
Independent	44.8	49.5	41.3	50.6	44.8	54.3	47.6
Constrained	44.5	51.2	42.3	53.5	45.6	55.6	48.8
CLAP (ours)	**45.5**	**52.0**	**44.8**	**54.7**	**48.2**	**56.9**	**50.4**
Supervised	60.7	66.4	61.4	68.6	63.2	69.7	65.0

Table 4.9: Ablation study comparing different contextual translation techniques for label projection. Performance is measured by downstream EAE performance.

Table 4.8. We see that CLAP performs the best with an average improvement of 2 F1 points over the next best baseline of Awesome-align and 6.5 F1 points over the zero-shot baseline. This result shows our model CLAP is a generalizable label projection technique and agnostic to the underlying translation model.

4.4.3 Ablation Study for CLAP

To study the impact of using instruction-tuned models for *contextual translation*, we conduct an ablation study comparing CLAP with the following strong baselines: (1) **Independent** translation uses the translation model $\mathcal{T}$ to independently (without any context of the input sentence) translate the source text labels to the target language (i.e. $\mathbf{y}^{tgt} = \mathcal{T}(\mathbf{y}^{src})$), (2) **Constrained** translation which uses a decoding constraint to carry out the faithfulness requirements. More specifically, during translation, it limits the generation vocabulary to the tokens in the translated sentence x^{tgt}. We follow De Cao et al. [30], Lu et al. [129] for implementing these constraints.

We extrinsically evaluate the model performances of the techniques on the task of EAE using the MMT translation model [5] and show the results in Table 4.9. We notice how simple independent translations can provide strong gains over the zero-shot model, but contextual translation can provide higher gains. The improvement of 1.6 F1 points of CLAP over

[5]Since decoding-time constraints for the Constrained model can't be applied to GMT

Lang	ha	ig	ny	rw	sn
Zero-shot	**72.9**	46.4	49.0	45.0	50.2
Awesome-align	72.2	**64.1**	**64.9**	**55.9**	55.4
EasyProject	72.0	54.6	50.5	54.5	42.5
CLAP (ours)	69.9	60.5	58.7	53.6	**59.7**
	sw	**xh**	**yo**	**zu**	**qu**
Zero-shot	**88.6**	61.0	**33.6**	**67.1**	37.9
Awesome-align	82.9	52.4	30.8	57.9	46.1
EasyProject	81.3	50.6	25.2	44.3	44.1
CLAP (ours)	80.7	**61.3**	30.6	54.4	**48.7**

Table 4.10: Extrinsic evaluation of the different label projection techniques using translate-train using GMT for NER for 10 low-resource languages.

the Constrained model highlights the significance of using an instruction-tuned model for contextual translation.

4.4.4 CLAP for Low-Resource Languages

To cater our model to a wide range of languages, we study the applicability of CLAP for low-resource languages. Specifically, we consider the task of NER for 10 low-resource languages from Africa and South America. For the test datasets, we utilize MasakhaNER [1] for 9 African languages: Hausa (ha), Igbo (ig), Chichewa (ny), Kinyarwanda (rw), chShona (sn), Kiswahili (sw), isiXhosa (xh), Yorùbá (yo), isiZulu (zu), and refer to Zevallos et al. [247] for the South American language Quechua (qu). We conduct an extrinsic evaluation of translate-train models trained on the English CoNLL training data[6] using the GMT model and present the results in Table 4.10. We observe that this is a particularly challenging setting as all the label projection techniques fail to improve over the zero-shot model for 4 languages. Our model CLAP performs decently, improving for 6 languages and performing the best for 3 languages. This result is particularly encouraging as our model uses a small and English-centric 13B Llama-2 model, and we hope these improvements will increase in

[6]For qu, we only use 3,000 CoNLL training data points due to budget constraints.

	EAE		NER			Avg
	ar	zh	it	es	id	
Zero-shot	36.3	47.3	79.4	74.5	53.1	58.1
Awesome-align	32.8	30.1	**77.5**	69.6	51.4	52.3
EasyProject	17.0	11.5	65.9	62.6	51.8	41.8
CLAP (ours)	**34.3**	**39.5**	73.4	**75.0**	**57.4**	**55.9**

Table 4.11: Extrinsic evaluation of the different label projection techniques in terms of downstream model performance using translate-test using GMT for EAE and NER. Avg = Average

future works by simply utilizing larger and multilingual LLMs.

4.4.5 CLAP for Translate-Test

Another popular technique for cross-lingual transfer is translate-test [79, 180] which was discussed in Section 4.1.1. As part of this analysis, we study the applicability of CLAP for translate-test using extrinsic evaluation on Arabic (ar) and Chinese (zh) for EAE and Italian (it), Spanish (es), and Indonesian (id) for NER. We show the results in Table 4.11. Overall, we see how CLAP outperforms both the other methods, significantly achieving the best scores for 4 out of the 5 languages. EasyProject performs the worst as it uses the translation model twice, causing higher error propagation. We also note how translate-test doesn't yield improvements over the zero-shot baseline, especially for EAE, as it requires label projection twice (once each for trigger and arguments), thus leading to error propagation.

4.5 Conclusion

In this chapter, our work proposes a novel approach, CLAP, for label projection, which utilizes contextual machine translation using instruction-tuned language models. Experiments on two structure prediction tasks of EAE and NER demonstrate the effectiveness of CLAP compared to other label projection techniques. Furthermore, intrinsic evaluation provides insights to

justify our model improvements. Overall, we lay the foundation for exploring the utilization of contextual translation, and future works can use it for various other applications as well.

Chapter 5

DiCoRe: Enhancing Zero-shot Event Detection via Divergent-Convergent LLM Reasoning

Training effective ED models requires large amounts of expert-annotated domain-specific data, which is highly costly and labor-intensive. This underlines the need to develop zero-shot systems that can perform ED robustly without using any training data. Although large language models (LLMs) have shown strong zero-shot performance across various tasks [154, 115]; their effectiveness on ED remains limited [57, 86], due to the requirement of extensive domain knowledge and the complex structural nature of ED.

ED requires deep reasoning and imposes several intertwined constraints: study of the large, closed event ontology and ensuring the event types must be chosen from it; semantic understanding of the input passage and precisely identifying domain-specific triggers within it; and conforming the output to a strict, machine-parsable structured format. Encoding these constraints as natural language instructions in the prompt overloads the LLM cognitively, making it harder to effectively apply its reasoning skills [205]. This increased difficulty in reasoning causes failures, such as missing relevant events, predicting irrelevant ones, and

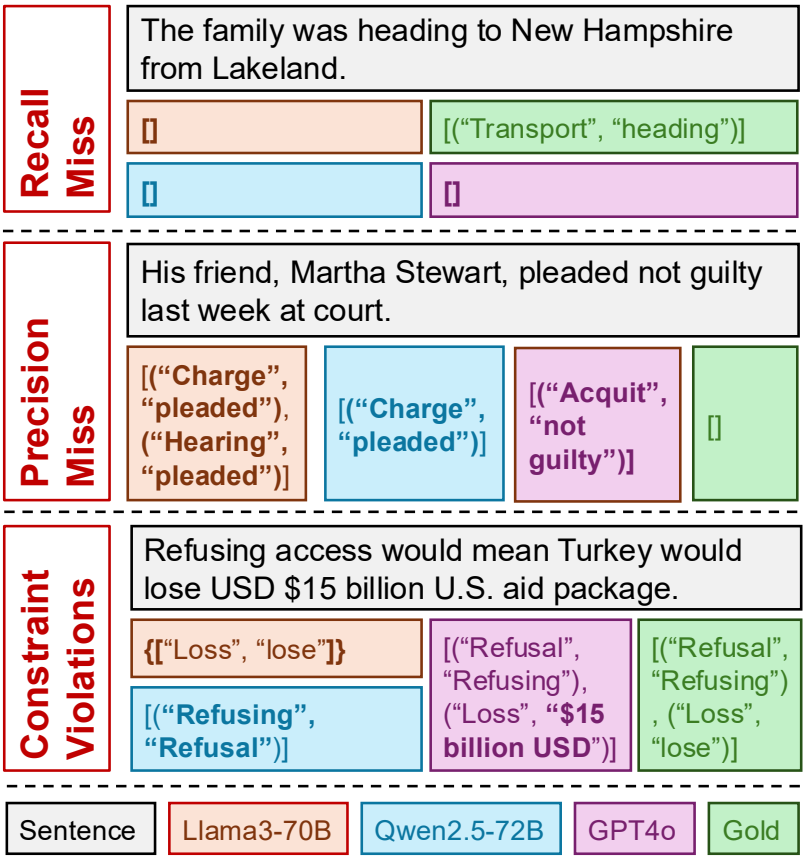

Figure 5.1: (top) Illustration of how prompting LLMs directly for Event Detection (ED) with all the task constraints can lead to precision, recall, and constraint violations (incorrect JSON, trigger not in sentence) across various LLMs. The errors are highlighted in **bold**.

struggling to follow the expected format, as shown in Figure 5.1.

To this end, we propose DiCoRe, a novel multi-agent pipeline introducing **Di**vergent-**Co**nvergent **Re**asoning, that facilitates better ED performance by reducing the cognitive burden of constraint adherence on the LLM. DiCoRe comprises two major agents in a pipeline: Dreamer and Grounder. (1) Dreamer fosters divergent reasoning by prompting in an unconstrained, open-ended manner. This encourages broad semantic exploration of potential event mentions by removing rigid task constraints and, in turn, boosts the recall. (2) Grounder introduces convergent reasoning by mapping Dreamer's free-form predictions to the task-specific closed event ontology. To alleviate the constraint adherence burden on the LLM, we employ a finite-state machine (FSM) to encode structural and task-specific

constraints. This FSM guides the generation process through constrained decoding, ensuring that the output adheres to the task requirements. Finally, we add an LLM-Judge to verify the grounded predictions against the original task instructions, ensuring high precision by filtering irrelevant predictions.

We conduct extensive experiments on six datasets from five domains across nine LLMs. Compared with various existing LLM inference works [57, 221, 165], we show how DiCoRe performs the best with average improvements of 4-5% F1 Trigger Classification and 5.5-6.5% F1 Event Identification over the best baselines. DiCoRe, without any training, also consistently improves over transfer-learning baselines [73, 183] fine-tuned on 15-30k datapoints by at least 5-12% F1. Furthermore, we demonstrate that DiCoRe provides 1-2% F1 gains while using 15-55x fewer inference tokens relative to strong thinking-based models and chain-of-thought (CoT), highlighting the significance of our proposed divergent-convergent reasoning.

In summary, we make the following contributions in this chapter: (1) We propose Dreamer, introducing divergent reasoning to improve event coverage. (2) We develop Grounder, performing convergent reasoning to align free-form predictions to the event ontology. (3) We design FSM-guided decoding to enforce task-specific structure during inference. Through extensive generalized evaluations across six datasets, five domains, and nine LLMs, we demonstrate the generalizability and efficacy of DiCoRe, establishing it as a robust zero-shot ED framework.

In terms of organization, we first discuss some relevant related works in Section 5.1 and provide the main multi-agent methodology of DiCoRe in Section 5.2. Then, we provide details about the extensive experimental setup and baselines in Section 5.3. Finally, we provide the major results and analyses in Section 5.4. The work in the chapter has been published as Parekh et al. [166].

5.1 Background and Related Work

Here, we discuss some specific related works discussing zero-shot ED frameworks and works in the direction of unconstraining LLMs for better reasoning.

Zero-shot Event Detection: To explore generalizability, initial works posed ED as a question-answering [42] or machine-reading comprehension problem [127]. Various works explored transfer and joint learning across various IE tasks to build more universal IE models [131, 51, 124]. Some works have explored posing ED as a generative text-to-text approach with event-based templates [130, 121, 73], even for zero-shot cross-lingual transfer [85, 162]. However, these works require source data training for zero-shot transfer, limiting their utility. Recent works have also explored the utility of zero-shot prompting with LLMs - concluding their sub-par performance [57, 113]. Other works have explored utilizing LLMs for data generation [134, 252, 165] to aid better generalizability. In our work, we focus on improving LLMs' zero-shot task generalizability to ED without any model fine-tuning.

Unconstraining LLMs for Better Reasoning: LLMs show immense language understanding and generation capabilities, but they need instructions and constraints to aid in meaningful human tasks [155]. However, imposing constraints also reduces LLM reasoning capabilities [205, 209, 15]. To this end, works have explored constrained decoding by altering the output probability distribution [227, 146, 248]. Some works explore grammar-aligned decoding strategies [58, 168]. However, such strict enforcement has been shown to hurt LLMs' generations. Recently, Tam et al. [205] explored better prompt design on math reasoning to unburden the constraints on the LLM. With similar inspiration, we explore decoupling LLMs from constraints to improve reasoning in our work.

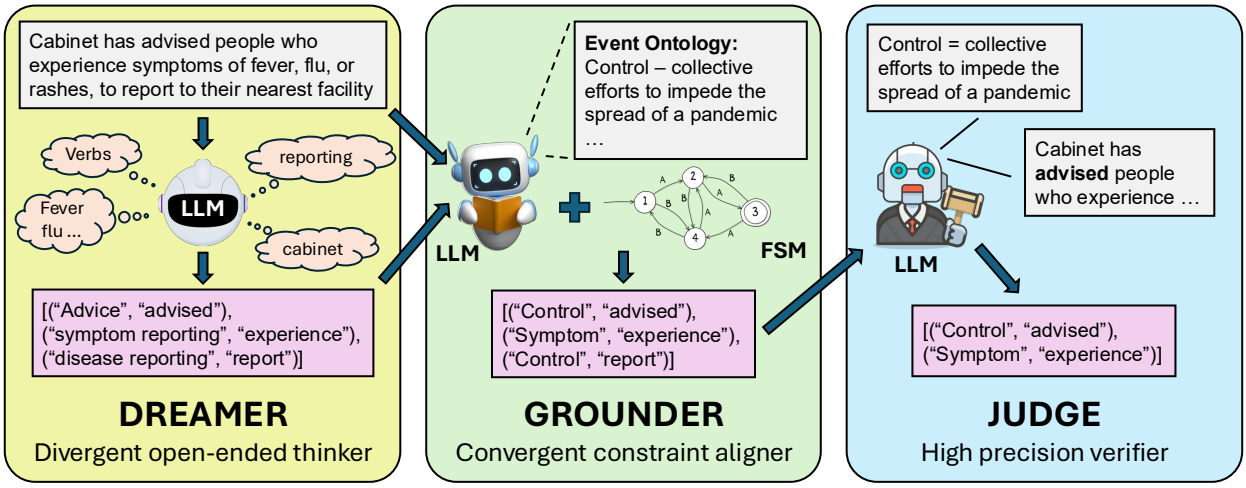

Figure 5.2: Illustration of our DiCoRe pipeline. First, the Dreamer reasons divergently in an open-ended, unconstrained manner about all potential events in the text and generates free-form event names. Next, the Grounder reads the event ontology and convergently grounds the free-form predictions to one of the event types. It uses a finite-state machine (FSM) guided constrained decoding to enforce task-specific constraints. Finally, the Judge evaluates each prediction and verifies its validity at a holistic scale.

5.2 Methodology

In our work, we frame ED through a generative outlook f_{gen} [160, 85] as they provide stronger zero-shot performance [73] and are better suited for LLMs. We consider a structured list of tuples as the output generation, as they provide stronger performance and are easy to parse [221]. However, these considerations introduce constraints (encoded as task instructions in the LLM prompt) like the predicted event is from the provided list, the predicted trigger phrase is in the input text, and the output generation follows the JSON format, as technically described below.

$$Y = f_{gen}(X; \mathcal{E}) \qquad \text{where}$$

$$Y = \text{“}[(e_1, t_1), ...(e_n, t_n)]\text{”} \tag{5.1}$$

$$t \in X \qquad \forall t \in \{t_1, ...t_n\} \tag{5.2}$$

$$e \in \mathcal{E} \qquad \forall e \in \{e_1, ...e_n\} \tag{5.3}$$

We argue that these structured constraints inherent to ED (Eq. 5.1-5.3) increase the cognitive load on LLMs, making reasoning more difficult [205]. This is one of the contributing factors to LLMs' subpar performance for ED [86]. To address this, we propose DiCore, a novel multi-agent pipeline that decouples and reduces constraint adherence by designing sub-agents to induce divergent open-ended discovery, convergent alignment, and constrained decoding. DiCoRe is lightweight, does not require additional training, and can be seamlessly applied to any LLM. Specifically, DiCoRe comprises a three-stage pipeline of a Dreamer-Grounder-Judge, as illustrated in Figure 5.2, and described below.

5.2.1 Dreamer

Our first agent, Dreamer *aka Divergent open-ended thinker*, is designed to promote open-ended divergent discovery, encouraging the LLM to achieve high recall by freely identifying potential events without being constrained by the predefined event ontology. Specifically, the Dreamer f_d removes the task-specific event constraint (Eq. 5.3), relaxes the trigger constraint (Eq. 5.2), and prompts the LLM to extract event mentions directly from the input sentence X as

$$Y_d = \text{“}[(e'_1, t_1), ...(e'_n, t_n)]\text{”} = f_d(X)$$

where each e'_i is a free-form LLM-generated natural language event name. Notably, e'_i need not adhere to the event ontology $\mathcal{E}$. We provide an illustration of the LLM prompt in Figure 5.3.

By removing explicit references to the event ontology, the instructions become less restrictive and more semantically intuitive for the LLM. This simplification enables the model to divergently reason on the underlying semantics of the text, rather than rigidly aligning with predefined categories. This open-ended setup encourages broader event discovery, improving recall by allowing the model to identify diverse or implicit event types. Though it may lower precision, it produces a rich candidate set for downstream refinement.

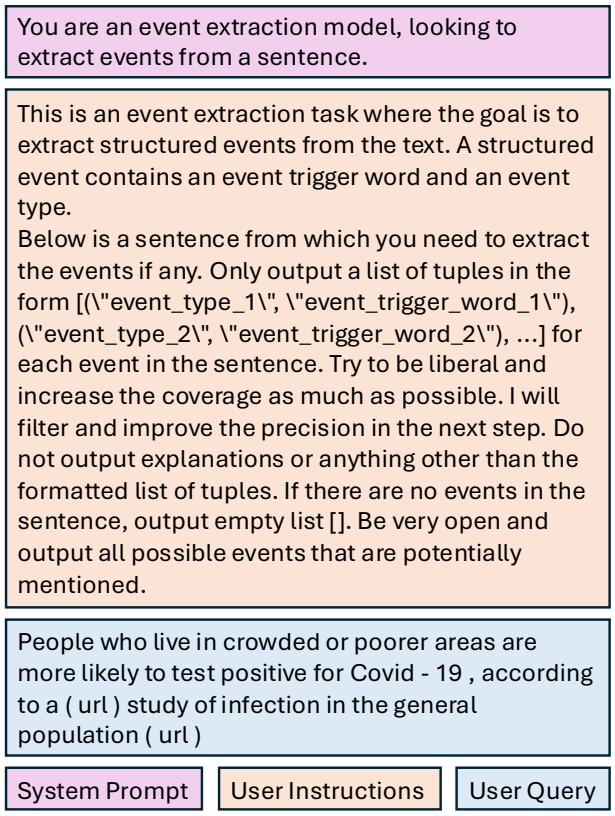

Figure 5.3: Illustration of the prompt utilized for the Dreamer agent. To encourage divergent thinking, we remove event-based constraints from the model instructions. Furthermore, we add sentences that encourage the model to be liberal and open in its predictions.

5.2.2 Grounder

The second agent, Grounder *aka Convergent constraint aligner*, convergently aligns the Dreamer's open-ended predictions Y_d with the closed, task-specific event ontology $\mathcal{E}$, while filtering the events that are not mappable. Technically, the Grounder f_g infuses the original task-specific constraints into the prompt to generate the grounded event mentions Y_g as

$$Y_g = \text{“}[(e_1, t_1), ...(e_m, t_m)]\text{”} = f_g(X; \mathcal{E}, Y_d)$$

An illustration of the Grounder prompt and expected output is shown in Figure 5.4.

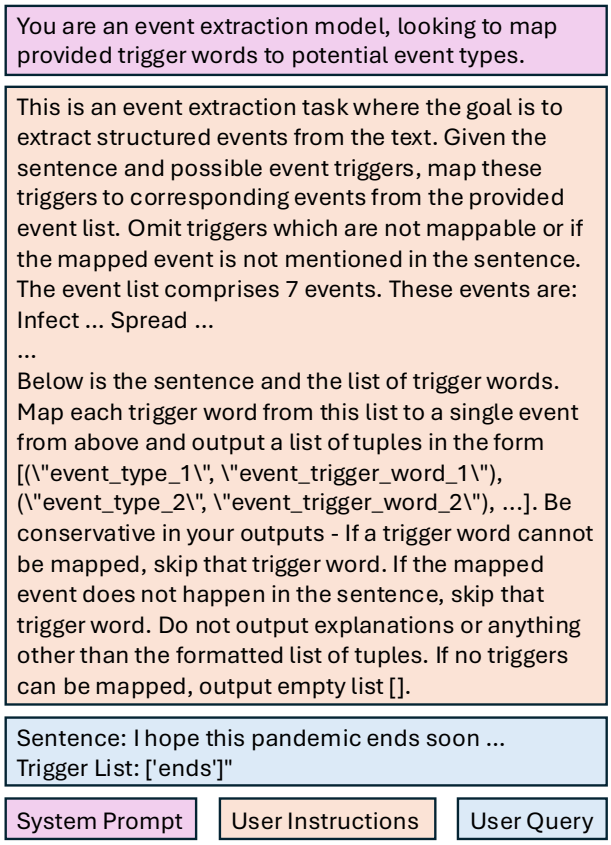

Figure 5.4: Illustration of the prompt utilized for the Grounder agent. To encourage convergent thinking and alignment, we add event-based constraints in the model instructions. Furthermore, we add sentences that encourage the model to be more conservative in its predictions.

FSM-guided decoding for constraint enforcement: To reduce the burden of constraint-following on the LLM and ensure strict adherence to the task format, we incorporate a constrained decoding mechanism guided by a finite-state machine (FSM). Inspired by recent work [227, 248], the FSM explicitly encodes structural and task-specific constraints (Eq. 5.1–5.3) within the decoding process. We construct and demonstrate an FSM to encode constraints for our ED task in Figure 5.5.

The FSM states represent decision points (e.g., whether the sentence contains an event, which event type $e \in \mathcal{E}$ to choose, which trigger $w \in X$ to assign, etc.), and the transitions denote valid LLM generations at each point (e.g., list of event types in $\mathcal{E}$, trigger words in the sentence). As shown in Figure 5.5, generation proceeds step by step: starting with the

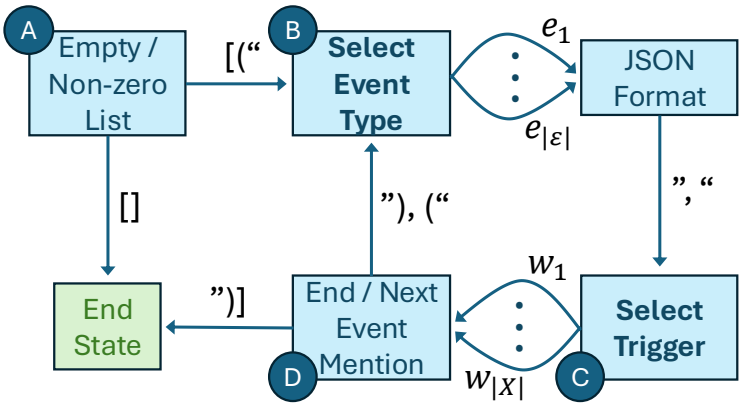

Figure 5.5: Finite state machine (FSM) illustration for guiding decoding to enforce constraints. Here $e_1, \ldots, e_{|\mathcal{E}|} \in \mathcal{E}$ represent all the possible event types and $w_1, \ldots, w_{|X|} \in X$ represent the atomized phrases in the sentence X.

event/no-event decision (state A), followed by selecting an event type (state B), then its trigger (state C), and finally deciding whether to generate another event mention or terminate (state D). To ensure that the generations are natural, the FSM states are partitioned in alignment with the LLM tokenizer, i.e., the states are chosen such that the sequence of transition tokens is the most probable tokenization of the output text Y_g.

We implement this FSM using the Outlines library [227] integrated into a vLLM inference framework [102]. The module takes as input the ontology, input sentence, LLM, and output JSON schema (potentially expressed as a grammar). Each FSM state transition is encoded as an Outlines `choices` list, thereby restricting the LLM's output vocabulary to only valid strings for that transition. For example, the set of possible event types or candidate trigger words is directly provided as the restricted vocabulary, and transitions with a single option are handled deterministically. The selected string then determines the next FSM state.

This design enforces structural validity during decoding: at each step, irrelevant tokens w.r.t. FSM transitions are zeroed out, ensuring the LLM can only generate ontology-compliant outputs. Our implementation currently supports JSON tuples of the form (event type, trigger), making it directly applicable to any ED dataset. More generally, because the transition and state mappings can be automatically constructed from the grammar of task constraints, the

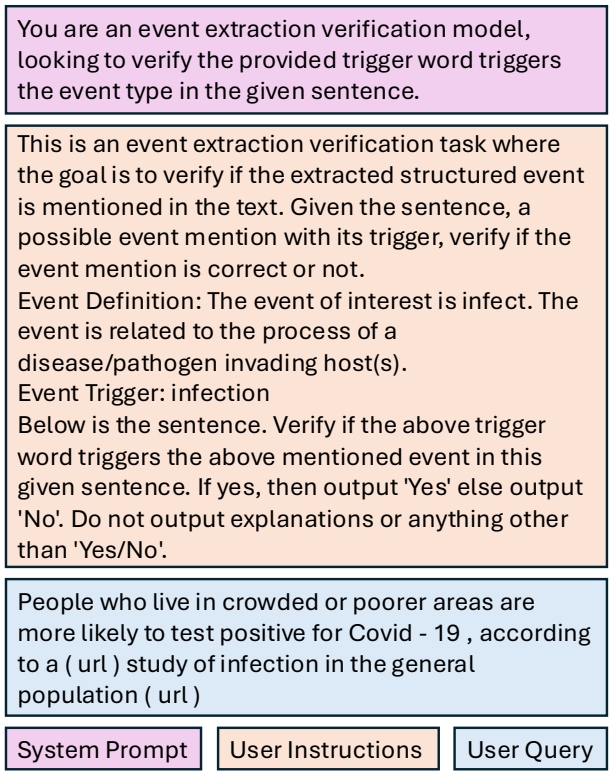

Figure 5.6: Illustration of the prompt utilized for Judge agent. To encourage convergent thinking and alignment, we add event-based constraints in the model instructions. Furthermore, we add sentences that encourage the model to be more conservative in its predictions.

approach can be adapted for other output formats and structured prediction tasks.

5.2.3 Judge

The final agent of our pipeline, Judge *aka High precision verifier*, serves to ensure each predicted event mention adheres to the original task instructions. Specifically, for each candidate pair (e_i, t_i), the Judge f_j evaluates the hypothesis that the trigger t_i expresses the event type e_i in the context of the input sentence X as

$$y_j^i = \text{“}Yes/No\text{”} = f_j(e_i, t_i, X; \mathcal{E})$$

All predictions with $y_j^i = \text{“}Yes\text{”}$ are accepted into the final output, while the others are rejected. We provide an illustration of the prompt in Figure 5.6.

Dataset	Domain	# Doc	# Event Mentions	Avg. Doc Length
MAVEN	General	250	623	24.5
FewEvent	General	250	250	30.5
ACE	News	250	71	13.2
GENIA	Biomedical	250	2472	251.3
SPEED	Epidemiology	250	258	32.4
CASIE	Cybersecurity	50	291	283.1

Table 5.1: Data Statistics of the various ED datasets used in the experimental setup for extensive evaluation of our proposed multi-agent framework DiCoRe.

This verification step plays a crucial role in ensuring the semantic validity and task alignment of predictions at a holistic level. By filtering out irrelevant or uncertain outputs, the Judge substantially improves the precision of the overall system without requiring additional supervision or training.

5.3 Experimental Setup

In this section, we briefly describe the experimental setup comprising the datasets, baselines, evaluation metrics, and implementation details.

Datasets: We benchmark our model across six ED datasets spanning five diverse domains, listed as: (1) MAVEN [220] and (2) FewEvent [32] from the general domain, (3) ACE [37] from the news domain, (4) GENIA [95], from the biomedical domain, (5) SPEED [164], from the epidemiological/social media domain, (6) CASIE [185], from the cybersecurity domain.

We provide statistics about the test splits of the different datasets in Table 5.1. To avoid any distributional biases, following TextEE [86], we uniformly sample 250 datapoints from the combined train-dev-test splits of each dataset for evaluation. Since CASIE is a smaller dataset, we only use 50 test samples for this dataset. The table highlights the domain diversity of the datasets covering common domains like news and general, while also focusing on technical domains like biomedical and epidemiology. The datasets also show variation in the density,

with ACE, FewEvent, and SPEED being sparse, with up to 1 event mention/sentence. On the other hand, MAVEN, CASIE, and GENIA are denser with 2.5-10 event mentions/passage. Finally, we also show the variation in token length, with ACE being the lowest with an average of 13 tokens, while GENIA and CASIE are longer with 250-280 average tokens per input document.

Baselines: We consider two major baselines, described below: (1) Multi-event Direct (MD) [57] directly prompts the LLM to provide the final output in a single pass, and (2) Multi-event Staged (MS) [165] decomposes the task into two stages, where the first stage identifies the event and the second stage extracts the corresponding triggers. We also compare with other works like: (3) Binary-event Direct (BD) [133, 122] prompts the LLM to do binary classification for each event, (4) Decompose-Enrich-Extract (DEE) [193] utilizes instruction enrichment with schema information for ED, (5) GuidelineEE (GEE) [198], similar to Code4Struct [221], converts ED into a code-generation problem using Python classes and instantiations, and (6) ChatIE [226] decomposes ED via multi-turn conversations. We ensure consistent, structured outputs for each baseline to maintain fair comparisons. Furthermore, we add the Judge component to each baseline, if not already present, to ensure robust benchmarking of DiCoRe.

Base LLMs: We use the following LLMs for our base experiments: Llama3-8B-Instruct and Llama3-70b-Instruct from the Llama3 family [43], and Qwen2.5-14B-Instruct; Qwen2.5-72B-Instruct from the Qwen2.5 [235] LLM family; and GPT3.5-turbo and GPT-4o [19, 153] from OpenAI.

Evaluation Metrics: Following Ahn [7], Parekh et al. [165] we report the F1 scores for the following three metrics: (1) Trigger Identification (TI) - correct identification of triggers, and (2) Event Identification (EI) - correct classification of event types, and (3) Trigger Classification (TC) - correct identification of the trigger-event pair (event mention). To maintain consistency with traditional span-based evaluations, we used string matching to

LLM	Prompt Style	MAVEN (168)			FewEvent (100)			ACE (33)			GENIA (9)			SPEED (7)			CASIE (5)			Average		
		TI	TC	EI	TI	TC	EI	TI	TC	EI	TI	TC	EI	TI	TC	EI	TI	TC	EI	TI	TC	EI
Llama3-8B	ChatIE	33.7	7.3	13.8	20.8	10.2	**27.6**	30.6	24.9	46.8	8.6	3.2	11.3	28.4	15.5	**43.3**	10.8	3.6	20.4	22.2	10.8	27.2
	GEE	19.1	1.9	6.8	11.7	5.9	14.0	30.0	21.3	27.4	25.4	15.8	26.7	**35.9**	27.7	38.7	11.5	9.2	45.8	22.3	13.6	26.6
	DEE	33.7	6.0	9.2	21.1	10.6	17.8	26.9	19.8	36.1	25.3	16.9	**32.5**	29.1	20.3	39.2	8.7	7.6	48.3	24.1	13.5	30.5
	BD	**54.5**	10.7	12.3	22.3	9.9	15.0	34.2	19.5	31.4	**28.1**	11.2	30.2	35.3	24.7	37.2	16.8	7.4	44.5	31.9	13.9	28.4
	MD	45.9	2.8	4.0	25.2	9.5	15.2	35.6	22.4	30.1	22.8	15.3	25.4	34.9	**27.8**	42.4	10.3	8.8	47.9	29.1	14.4	27.5
	MS	46.2	10.3	11.2	20.2	10.2	17.0	26.7	17.6	23.1	27.6	**19.7**	30.5	34.1	27.3	40.6	11.9	10.3	48.3	27.8	15.9	28.4
	DiCoRe	53.5	**14.4**	**17.4**	**26.1**	**15.7**	25.0	**40.3**	**36.3**	**47.9**	25.8	15.4	30.0	35.5	23.6	42.4	**18.5**	**16.8**	**58.8**	**33.3**	**20.4**	**36.9**
Llama3-70B	ChatIE	47.9	19.8	24.8	33.3	20.8	**40.6**	45.5	37.9	47.0	14.6	6.4	17.3	41.8	31.0	50.9	12.9	10.2	48.9	32.7	21.0	38.3
	GEE	28.3	15.7	17.5	26.2	16.3	31.1	47.0	42.3	52.2	32.5	24.2	38.5	43.7	34.7	46.0	11.1	10.7	43.2	31.5	24.0	38.1
	DEE	60.8	14.8	16.4	34.0	21.3	33.6	47.4	38.3	45.4	39.2	30.5	46.0	41.7	32.2	44.7	16.6	16.4	63.1	40.0	25.6	41.5
	BD	63.0	13.9	15.2	34.0	14.5	22.6	49.1	36.6	41.7	**39.4**	26.5	45.4	**49.2**	33.6	45.7	16.5	11.7	48.8	41.9	22.8	36.6
	MD	**63.5**	14.2	14.7	34.0	20.9	32.6	51.2	40.2	46.8	36.8	28.9	43.0	45.4	36.8	49.0	13.9	13.7	64.4	40.8	25.8	41.8
	MS	33.9	21.6	22.3	35.3	24.9	39.9	49.9	42.8	46.9	37.4	**31.0**	45.0	43.8	35.5	49.6	14.0	14.0	59.5	35.7	28.3	43.9
	DiCoRe	62.5	**27.8**	**30.6**	**40.4**	**25.1**	36.1	**57.2**	**49.5**	**55.1**	38.6	**31.0**	**48.5**	45.0	**36.5**	**51.8**	**17.3**	**16.6**	**66.6**	**43.5**	**32.8**	**48.1**

Table 5.2: Main results comparing the zero-shot ED performance of our proposed DiCoRe with all other baselines for the Llama3-8B-Instruct and Llama3-70B-Instruct LLMs. TI: Trigger Identification, TC: Trigger Classification, EI: Event Identification. **bold** = best performance. (XX) = number of distinct event types.

map the generated outputs to input spans.

Implementation Details: We use TextEE [86] for our benchmarking, datasets, and evaluation setup. To restrict LLM's generation choices for the FSM-guided constrained decoding, we utilize Outlines [227] over vLLM inference [102]. We use Curator [136] to query the GPT family of LLMs. We deploy a temperature of 0.4 and a top-p of 0.9 for decoding. We report the averaged results over three runs for robust benchmarking.

5.4 Results and Analysis

In this section, we provide our main results and findings, and later provide supporting evidence through our analyses.

5.4.1 Main Results

We present the main zero-shot results for all baselines on the six datasets for Llama3 LLMs in Table 5.2. As seen from the average results (last three columns), DiCoRe performs the best, surpassing the best baseline of multi-event staged (MS) by a significant margin

LLM	Prompt Style	MAVEN (168)			FewEvent (100)			ACE (33)			GENIA (9)			SPEED (7)			CASIE (5)			Average		
		TI	TC	EI	TI	TC	EI	TI	TC	EI	TI	TC	EI	TI	TC	EI	TI	TC	EI	TI	TC	EI
Qwen2.5-14B	MD	53.0	17.6	20.9	28.8	**21.1**	**34.2**	28.3	24.5	42.1	24.8	18.8	26.7	37.7	33.0	**51.2**	15.8	15.8	**61.5**	31.4	21.8	39.5
	MS	46.5	20.8	24.6	24.8	18.9	32.1	33.6	26.3	32.5	25.4	19.2	27.7	38.9	34.3	46.1	16.3	16.1	54.5	30.9	22.6	36.2
	DiCoRe	**53.1**	**23.3**	**27.6**	**29.7**	19.3	30.4	**38.4**	**37.7**	**48.8**	**29.9**	**22.6**	**38.6**	**42.9**	**35.3**	46.5	**19.7**	**19.5**	58.8	**35.8**	**26.1**	**41.8**
Qwen2.5-72B	MD	49.4	21.6	24.1	17.0	12.3	21.0	28.8	25.8	30.3	30.5	27.0	36.3	**41.4**	**37.4**	**45.4**	11.0	10.4	57.9	29.7	22.4	35.8
	MS	39.9	23.6	25.4	25.0	21.0	**34.2**	42.5	40.4	42.5	26.7	23.6	34.1	40.6	35.5	45.2	10.5	10.5	49.1	30.9	25.8	38.4
	DiCoRe	**54.1**	**27.5**	**30.2**	**30.8**	**22.3**	32.9	**46.8**	**44.8**	**47.8**	**33.6**	**29.8**	**43.9**	40.6	34.7	41.4	**15.9**	**15.8**	**59.3**	**37.0**	**29.2**	**42.6**
GPT3.5-turbo	MD	**50.9**	17.4	20.4	23.2	14.6	27.0	40.9	36.2	42.5	**27.0**	**19.9**	31.4	**36.5**	**30.6**	41.8	10.0	9.9	51.1	**31.4**	21.4	35.7
	MS	48.2	15.5	17.2	23.7	**15.9**	29.8	40.7	37.4	42.3	23.2	19.0	26.3	33.0	23.7	35.5	7.7	7.1	44.4	29.4	19.8	32.6
	DiCoRe	48.1	**21.6**	**26.1**	**25.3**	15.6	**31.1**	**41.7**	**41.7**	**48.9**	26.2	19.5	**36.3**	32.4	27.2	**49.0**	**11.4**	**10.6**	**55.7**	30.9	**22.7**	**41.2**
GPT4o	MD	**61.8**	28.9	31.9	30.6	23.9	35.4	52.3	52.3	52.3	**41.0**	**36.5**	49.5	44.1	40.2	48.0	10.1	10.1	55.7	40.0	32.0	45.5
	MS	49.4	30.8	33.3	25.6	20.6	32.2	36.2	36.2	38.3	36.6	33.2	45.0	**45.7**	**40.4**	**50.1**	13.4	13.4	46.9	34.5	29.1	41.0
	DiCoRe	58.5	**32.2**	**35.6**	**36.1**	**28.4**	**38.5**	**54.9**	**54.9**	**56.6**	40.7	35.4	**51.2**	43.3	37.3	46.1	**16.7**	**16.7**	**58.8**	**41.7**	**34.2**	**47.8**

Table 5.3: Generalization results for zero-shot ED performance comparing DiCoRe with the best baselines for four other LLMs of Qwen2.5-14B-Instruct, Qwen2.5-72B-Instruct, GPT3.5-turbo, and GPT4o. **bold** = best performance. (XX) = number of distinct event types.

of 5.5-8% TI, 4-8.5% EI, and 4-5% TC. The performance disparity across different task decomposition methods of ChatIE, MS, and DiCoRe highlights how our divergent-convergent decomposition of Dreamer-Grounder provides a stronger inductive bias. Other baselines perform relatively better on datasets like GENIA/SPEED, as these are simpler datasets with fewer event types, thus requiring less cognitive reasoning. However, on the high-event datasets like MAVEN/FewEvent/ACE, which require more complex reasoning, DiCoRe, with its divergent-convergent reasoning, shows more significant improvement over the baselines.

Generalization across LLMs: To demonstrate the generalizability of DiCoRe, we benchmark it with the top-performing baselines on four additional LLMs from the Qwen and GPT families and show our results in Table 5.3. We note how DiCoRe performs the best across all LLMs with an overall average improvement of 5.5% TI, 6.5% EI, 4% TC over the multievent-staged baseline and 3.3% TI, 5.4%, 4.6% TC over the multievent-direct baseline. Across different LLMs, we note the strongest performance on GPT4o, followed by Llama3-70B-Instruct and Qwen2.5-72B, indicating how more parameters help better reasoning with DiCoRe.

Model Setting	Average F1		
	TI	TC	EI
Test on GENIA, SPEED, CASIE			
GOLLIE-7B	6.0	5.3	15.3
GOLLIE-34B	15.6	11.7	29.4
Llama3-8B DiCoRe	26.6	18.6	43.7
Llama3-70B DiCoRe	33.6	28.0	55.6
Test on all but ACE dataset			
ACE-trained DEGREE	20.9	11.0	21.3
Llama3-8B DiCoRe	31.9	17.2	34.7
Llama3-70B DiCoRe	40.8	27.4	46.7
Test on all but MAVEN dataset			
MAVEN-trained DEGREE	31.8	25.0	38.6
Llama3-8B DiCoRe	29.2	21.6	40.8
Llama3-70B DiCoRe	39.7	31.7	51.6

Table 5.4: Comparison of pure zero-shot DiCoRe with fine-tuned transfer-learning baselines. Underline indicates scenarios of DiCoRe improvements.

5.4.2 Comparison with Fine-tuned Transfer-learning Methods

Various works utilize transfer-learning and universal Information Extraction (IE) training for zero-shot cross-dataset ED [20, 124]. These works train on selected IE datasets and show performance on unseen IE datasets. We provide a comparison of DiCoRe with two such transfer-learning approaches: (1) DEGREE [73], a generative framework utilizing text-based event templates to generalize, (2) GOLLIE [183], a universal IE framework, fine-tuning LLMs on various IE instruction datasets. For DEGREE, we consider two versions where the source data is ACE and MAVEN, respectively. For GOLLIE, we consider the fine-tuned GOLLIE-7B and GOLLIE-34B models. We provide the averaged results across target datasets (not included in the source data) in Table 5.4. Through these results, we demonstrate how, despite no fine-tuning, DiCoRe consistently outperforms the fine-tuned transfer-learning baselines across all settings. On average, DiCoRe improves by 3-10% F1 using Llama3-8B-Instruct and 10-22% F1 using Llama3-70B-Instruct and GPT4o.

Agent	TI			TC		
	P	R	F	P	R	F
Llama3-8B-Instruct						
Dreamer	8.5	**64.3**	15.0	0.0	0.0	0.0
+ Grounder	20.4	47.9	28.6	15.5	37.1	21.9
+ FSM Decoding	22.3	56.8	32.1	16.2	**42.3**	23.4
+ Judge	41.8	39.0	**40.3**	**37.5**	35.2	**36.3**
MD Baseline	**48.4**	28.2	35.6	30.2	17.8	22.4
MS Baseline	22.0	33.8	26.7	14.4	22.5	17.6
Llama3-70B-Instruct						
Dreamer	15.5	**77.5**	25.8	0.0	0.0	0.0
+ Grounder	28.6	65.7	40.4	22.5	53.4	31.8
+ FSM Decoding	32.3	66.7	43.5	26.2	54.0	35.3
+ Judge	52.8	62.5	**57.2**	45.7	**54.0**	**49.5**
MD Baseline	57.2	46.5	51.2	44.0	37.1	40.2
MS Baseline	**66.4**	39.9	49.9	**57.0**	34.3	42.8

Table 5.5: Ablation Study on the ACE dataset highlighting the significance and contribution of each component of DiCoRe. P: Precision, R: Recall, F: F1 score.

5.4.3 Comparison with Reasoning baselines

Reasoning by verbalizing thoughts using additional tokens has commonly helped improve performance across a wide range of tasks [98, 104]. We evaluate the utility of reasoning, specifically Chain-of-thought (CoT) [224], along with thinking-based models like Deepseek-R1-Distilled-Qwen-32B (DS-Qwen-32B), Deepseek-R1-Distilled-Llama3-70B (DS-Llama3-70B) [31] and O1-mini [89] on our task of zero-shot ED in Table 5.7. We demonstrate how the baselines improve with additional reasoning; however, DiCoRe with the base models (Llama3-70B) consistently outperforms all these reasoning baselines (even O1-mini) while using 15-55x fewer tokens on average. We also show how our method is complementary to reasoning by demonstrating further improvements up to 1-2% F1 using reasoning with DiCoRe.

Sentence	Best Baseline Prediction	Dreamer Prediction	Grounder Prediction	Judge Prediction
cass apd ra gave birth to her first daughter.	[("**Life:Be-Born**", "**gave**")]	[("Birth", "gave"), ("Birth", "birth")]	[("Life:Be-Born", "birth")]	[("**Life:Be-Born**", "**birth**")]
After passing the island, the hurricane turned to the northeast, and became extratropical on September 8, before dissipating two days later.	[("**Change**", "**turned**"), ("**Change**", "**became**"), ("**Dissipating**", "**dissipating**")]	[("Movement", "turned"), ("Transition", "became"), ("Dissipation", "dissipating")]	[("Change_time", "turned"), ("Becoming_a_member", "became"), ("Dispersal", "dissipating")]	[("**Dispersal**", "**dissipating**")]
Covid-19 has led to social distancing, but we can still be together through the quarantine with online gaming!	[]	[("Social_Distancing", "distancing"), ("Quarantine", "quarantine"), ("Gaming", "gaming")]	[("prevent", "distancing"), ("control", "quarantine")]	[("**prevent**", "**distancing**"), ("**control**", "**quarantine**")]

Table 5.6: Qualitative examples comparing DiCoRe's predictions (per component) with the best baseline. We highlight the correct predictions in **green** and incorrect ones in **red**.

5.4.4 Ablation Study

To demonstrate the role of each component of our pipeline, we ablate and show the model performance as we add each component in DiCoRe for the ACE dataset for Llama3-8B and Llama3-70B LLMs in Table 5.5. For reference, we also show the precision/recall splits of the baselines. Dreamer achieves a high recall for TI (albeit a low precision) - demonstrating the utility of divergent unconstrained reasoning. Grounder helps align the predictions, causing a slight drop in recall while improving the precision. FSM Decoding helps largely improve the recall for Llama3-8B-Instruct by improving the mapping, and precision for Llama3-70B-Instruct by fixing any constraint violations. Finally, Judge largely boosts the precision of the model. Analysis of the baselines reveals that they are conservative, making a low number of high-precision predictions. In comparison, DiCoRe makes many more predictions, largely improving recall while maintaining reasonably high precision.

Base LLM	Prompt Style	Average F1		
		TI	TC	EI
Chain-of-thought Baselines				
Llama3-8B	MD + CoT	25.0	13.5	27.1
Llama3-8B	MS + CoT	28.4	17.6	31.9
Llama3-70B	MD + CoT	41.0	30.9	48.0
Llama3-70B	MS + CoT	40.5	31.6	47.1
Qwen2.5-72B	MD + CoT	34.9	27.1	43.6
Qwen2.5-72B	MS + CoT	36.2	28.8	40.8
Thinking-based model Baselines				
DS-Qwen-32B	MD	39.2	30.0	46.3
DS-Qwen-32B	MS	39.5	30.4	45.2
DS-Llama3-70B	MD	29.0	23.3	36.1
DS-Llama3-70B	MS	33.3	27.0	37.8
O1-mini	MD	40.2	32.5	44.7
DiCoRe base model results				
Llama3-8B	DiCoRe	33.3	20.4	36.9
Llama3-70B	DiCoRe	43.5	32.8	48.1
Qwen2.5-72B	DiCoRe	37.0	29.2	42.6
GPT4o	DiCoRe	41.7	34.2	47.8
DiCoRe improvements with reasoning				
Llama3-8B	DiCoRe + CoT	33.1	21.1	36.2
Llama3-70B	DiCoRe + CoT	43.0	33.1	**49.8**
Qwen2.5-72B	DiCoRe + CoT	37.0	29.1	43.5
DS-Qwen-32B	DiCoRe	**43.1**	**33.3**	49.5
DS-Llama3-70B	DiCoRe	41.4	33.0	48.3

Table 5.7: Comparison of DiCoRe with reasoning-based baselines like Chain-of-thought (CoT) and thinking-based models. Underline indicates DiCoRe improvements over baselines.

Qualitative Study: We provide some qualitative examples for each component of DiCoRe, while comparing the best baseline across the datasets in Table 5.6. We see how the best baseline often reasons incorrectly, leading to precision loss, or remains conservative, predicting nothing, leading to recall errors. The split across the three components shows how Dreamer generates many plausible event mentions, Grounder aligns and removes some, while Judge verifies and filters irrelevant ones. These examples provide the internal workings of DiCoRe, highlighting the significance of divergent-convergent reasoning.

5.5 Conclusion

In our chapter, we introduce DiCoRe, a novel multi-agent divergent-convergent reasoning pipeline of Dreamer-Grounder-Judge, aimed at decoupling the LLM from task-specific constraints, and indirectly better exploiting LLMs' reasoning. Through experimentation on six ED datasets from five domains across nine LLMs, we confirm how DiCoRe provides a stronger inductive bias, improving over other zero-shot baselines, fine-tuned transfer learning methods, and reasoning-focused approaches.

Chapter 6

SNaRe  Domain-aware Data Generation for Low-Resource Event Detection

ED frameworks are generally data-hungry, requiring costly, labor-intensive expert annotations. While zero-shot inference models have been developed [166], their performance is subpar relative to fully trained supervised models [57, 86]. To this end, synthetic data generation [231] (i.e., generating sentence and event annotations) has emerged as a promising alternative, particularly for practical use-cases in specialized domains.

However, existing generation approaches often focus on general-domain settings and fail to address the distinct challenges of specialized domains [197]. Weak supervision methods [68, 25] that use LLMs to generate labels for unlabeled sentences frequently introduce *label noise* (Figure 6.1(a)), where incorrect or incomplete labels arise due to weak LLM reasoning [86] or limited domain knowledge [197]. Downstream training on such incorrect labels can cause spurious bias propagation. Conversely, recent generation approaches [92, 134] that utilize LLMs' self-knowledge to jointly generate labels and sentences struggle with *domain drift* (Figure 6.1(b)), often synthesizing sentences that are misaligned with the target domain.

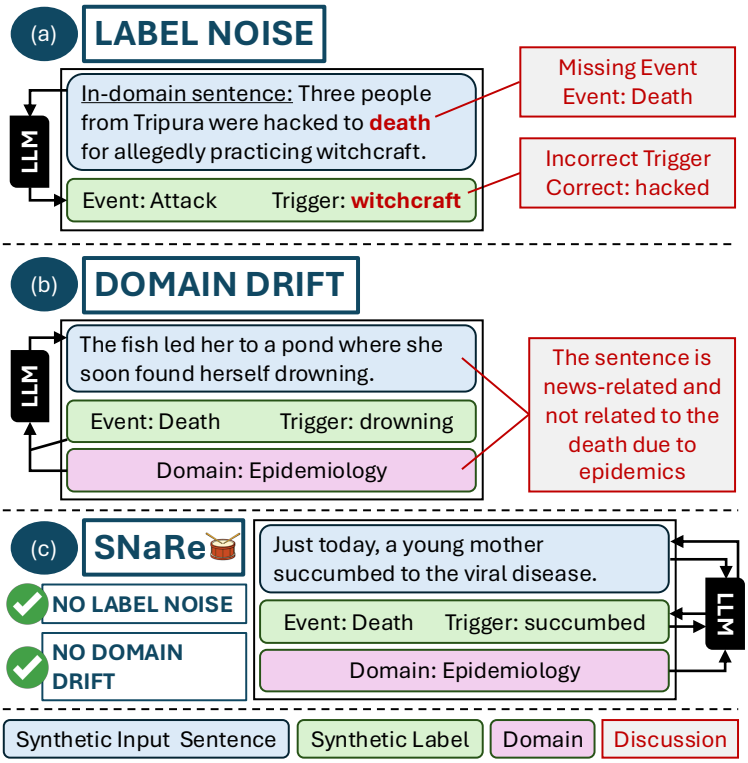

Figure 6.1: Highlighting the errors of existing data generation approaches. (a) Using LLMs to generate the labels from in-domain sentences leads to *label noise* owing to poor LLM reasoning. (b) Utilizing LLMs to generate sentences conditioned on event and domain causes *domain drift*, wherein the synthetic sentence is not aligned with the target domain. Finally, in (c), we illustrate how SNaRe minimizes both errors to generate higher-quality synthetic data.

This can be attributed to the lack of utilization of target domain information for generation and can drastically hamper model training as lexical and structural cues are highly influential for ED [212]. Overall, *label noise* and *domain drift* reduce synthetic data quality, eventually leading to subpar supervised downstream model performance.

To this end, we propose SNaRe, a novel domain-aware, three-stage data synthesis LLM pipeline comprising the **S**cout, **Na**rrator, and **Re**finer agents. Scout surveys unlabeled target domain data to identify salient triggers via prompt-based trigger extraction. Using corpus-level statistics for automated aggregation and filtering, Scout curates a list of high-quality domain-specific triggers per event type. Next, Narrator samples from these domain-specific triggers and utilizes LLMs to synthesize diverse sentences for each event type. Utilizing specialized

and diverse domain information for conditional generation aids Narrator in generating more domain-aligned sentences, eventually reducing domain drift in the synthesized sentences. Since Narrator sentences could mention additional events apart from the input set, we design the Refiner to utilize LLM inference to annotate such missing events in these sentences. Narrator's conditional text generation and Refiner's missing label annotation aid in reducing the label noise and ensuring high data quality. We provide an illustration of SNaRe's generation in Figure 6.1(c).

We benchmark SNaRe on ED datasets from three domains: ACE [37] (news), SPEED [164] (epidemiology), and GENIA2011 [95] (biomedical). For evaluation, we report the ED performance of DEGREE [73] trained on the synthesized data. Across the zero-shot and few-shot settings, SNaRe performs the best, outperforming the previous state-of-the-art baselines [36, 134] by an average of 3-7% F1 points. Under multilingual generation for Arabic and Chinese, SNaRe outshines even more, with improvements of 4-20% F1 over the best baseline. Our analysis reveals how SNaRe's synthesized triggers overlap 4-11% more (relative to baselines) with the gold trigger set, demonstrating the reduction in domain drift. Finally, human evaluation provides qualitative evidence for SNaRe's superior data annotation quality and domain alignment.

We start with discussion of relevant related works in Section 6.1. Next, we provide our proposed multi-agent data synthesis pipeline SNaRe in Section 6.2. Then, we describe our extensive experimental setup in Section 5.3, followed by results and extensive analyses in Section 5.4. This chapter's work has been published as Parekh et al. [165].

6.1 Background and Related Work

Here, we discuss specific related works discussing data generation paradigms for information extraction and low-resource event detection in general.

Data Generation for Information Extraction LLM-powered synthetic data generation has been successful for various NLP tasks [123, 222, 229, 190]. For information extraction, works have explored knowledge retrieval [22, 9], translation [162, 105], data re-editing [106, 80], and label extension [252]. Recent works utilize LLMs to generate labels for sentences [25, 245, 217, 206], while some other works explore the generation of sentences from labels [92, 134]. Our work introduces SNaRe, focused on infusing better domain-specific information for data generation.

Low-resource Event Detection To address the growing need for event detection across expanding domains, prior works have explored transfer learning via Abstract Meaning Representation [87], Semantic Role Labeling [250], and Question Answering [133]. Reformulating ED as a conditional generation task has also aided low-resource training [73, 76, 85]. Recently, LLM-based reasoning [112, 56, 221, 166] and transfer-learning [20] have been explored, but their performance remains inferior to supervised models [86]. This motivates efforts in LLM-powered synthetic data generation for low-resource ED.

6.2 Methodology

In this chapter, we focus on domain-aware synthetic data generation using LLMs to alleviate the need for expert-annotated training data. By generating a large dataset $D_s = \{(X, Y)\}$, we can train domain-specific downstream ED models using minimal supervised data.

Existing weak-supervision approaches [141, 223] utilizing automatic methods to assign labels to unlabeled sentences often generate incorrect labels (*label noise*). This can be attributed to the domain-specific context understanding and deep reasoning requirement of ED, leading to poor automatic label quality, even when using recent LLMs [86]. Another route of approaches that utilize LLMs to generate sentences conditioned on labels [187, 92], i.e., synthesize X for a designated Y, often curate sentences that are distributionally divergent from the target domain (*domain drift*). This is mainly because these approaches focus on the

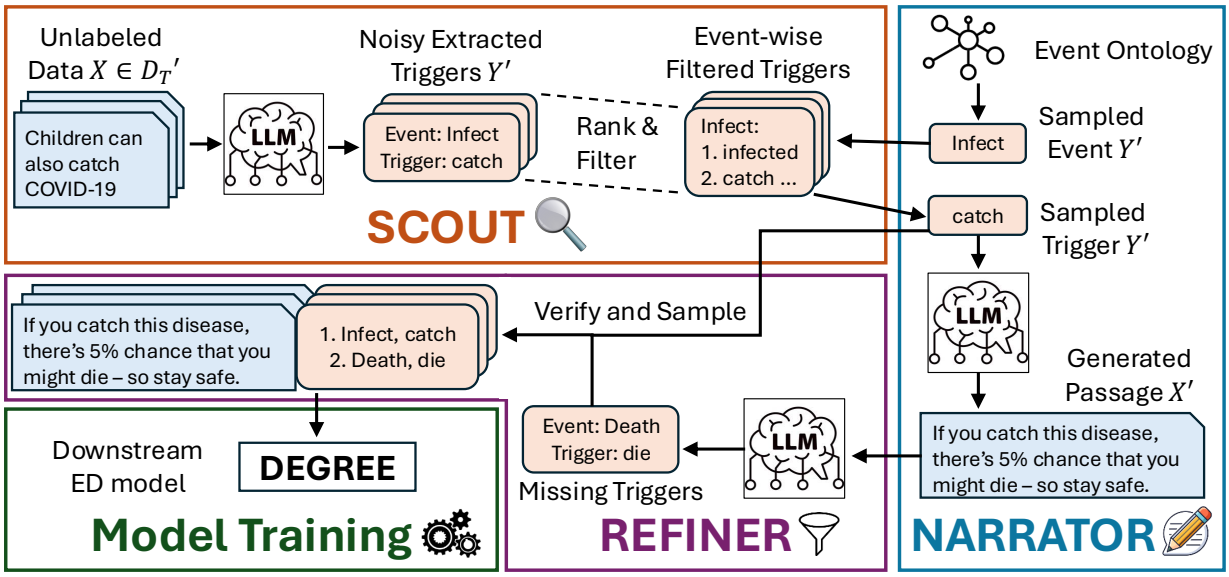

Figure 6.2: Model Architecture Diagram highlighting the various agents of SNaRe. First, Scout extracts and filters domain-specific triggers, then Narrator generates passages conditioned on these triggers. Finally, Refiner adds any missing annotations and sample N data points per event type for downstream training.

general domain and fail to utilize any target domain signals in their generation. Both label noise and domain drift hurt the synthetic data quality, in turn, diminishing the downstream supervised model performance.

To mitigate these issues, we propose SNaRe, a domain-aware multi-agent data synthesizer, that generate and verifies LLM generations to ensure high-quality data [77], comprising three agents: **S**cout, **Na**rrator, and **Re**finer: Scout studies unlabeled target domain data D_T' to curate domain-specific triggers, in turn, reducing domain drift. Narrator generates domain-specific sentences conditioned on Scout's curated triggers, while Refiner adds additional annotations to ensure high-quality labels. Overall, SNaRe is a training-free LLM inference pipeline that is easily deployable and scalable across domains. We provide our architectural diagram in Figure 6.2 and explain each agent of our pipeline below.

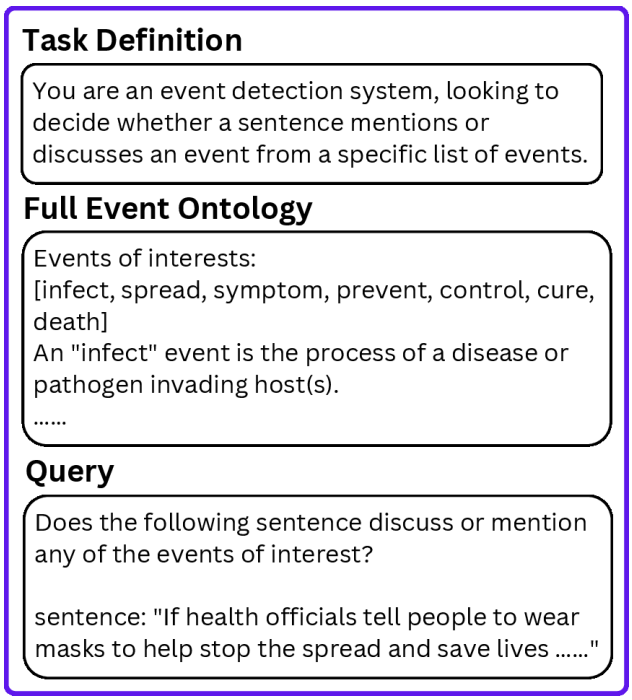

Figure 6.3: Prompt setup for stage 1 of Event Classification for the Scout agent.

6.2.1 Scout

Scout is tasked with the curation of domain-specific triggers that are later utilized for sentence generation. Events can assume a wide range of triggers depending on the domain and context. For example, an "Attack" event can be triggered by *war*, *killed* in news, or *breach*, *phish* in cybersecurity, or *infect*, *transmit* in epidemiology domains.

Thus, unlike past works [134] that utilize only LLMs' internal knowledge for trigger generation, we develop **Scout**, which extracts high-precision domain-specific triggers using unlabeled target domain data D'_T. Specifically, trigger extraction involves a two-stage prompt setup: (1) The first stage is tasked with identifying and filtering possible event types mentioned in the target domain sentence, and (2) The second stage aims to find the most appropriate trigger word from the unlabeled sentence for each filtered event type. We provide the prompts for Scout in Figure 6.3 and 6.4.

To ensure high precision of the triggers, we develop an aggregation and filtering mechanism by incorporating corpus-level statistics. Specifically, for each event type, we aggregate the

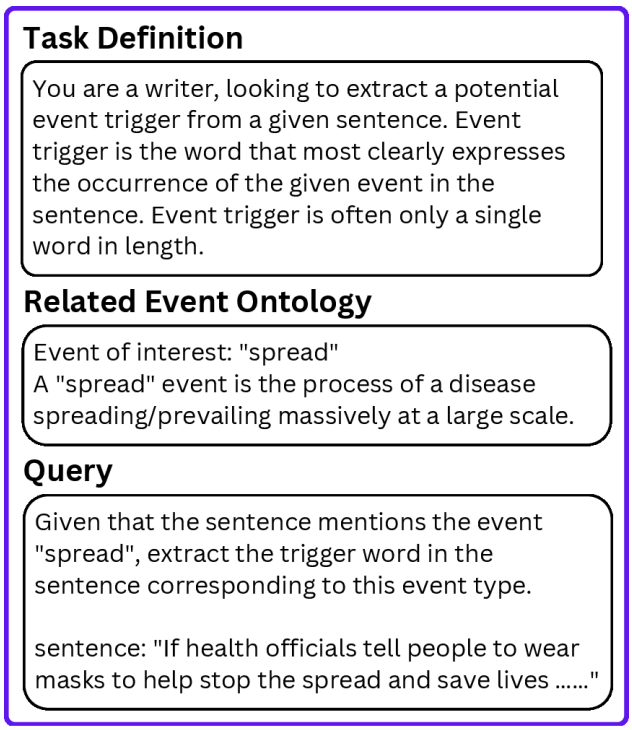

Figure 6.4: Prompt setup for stage 2 of Trigger Extraction of the Scout agent.

counts of the triggers at the corpus level and filter out the top $t = 10$ triggers as the curated list of high-quality domain-specific event-indexed triggers. These triggers carry important target domain signals that help in generating domain-specific sentences (§ 6.2.2), in turn reducing domain drift of our synthetic data.

6.2.2 Narrator

Narrator is tasked with the synthesis of domain-specific sentences for our synthetic dataset. Existing works [92, 134] do not utilize any target domain information, which causes domain drift in their synthesized sentences. Instead, in our work, we condition the Narrator to utilize the rich and diverse domain-specific triggers from Scout to synthesize domain-specific sentences, which, in turn, reduce domain drift.

Specifically, Narrator samples 1-2 event types per synthetic data instance and corresponding domain-specific triggers from Scout's curated trigger list - constituting the label Y. Next, it prompts the LLM with the task instructions and the event definitions, and asks it to

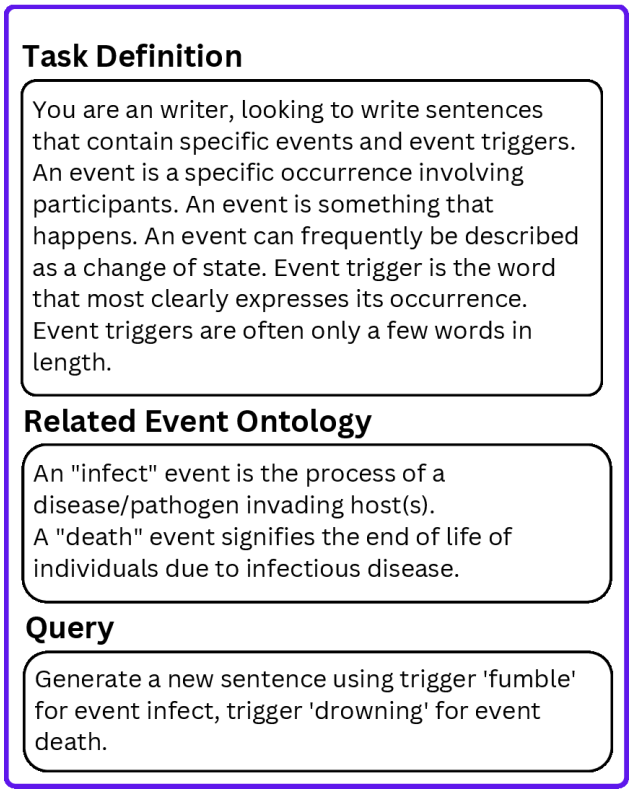

Figure 6.5: Prompt setup for conditional data generation for the Narrator agent.

generate a passage X' that mentions the sampled events using the sampled triggers (Y). We illustrate this prompt in Figure 6.5. The generated sentences are naturally more aligned with the target domain owing to the conditioning on the domain-specific triggers (qualitative examples shown in § 6.4.8). Further domain-specific adaptations are possible by fine-tuning the LLM on unlabeled domain-specific data (analysis in § 6.4.7).

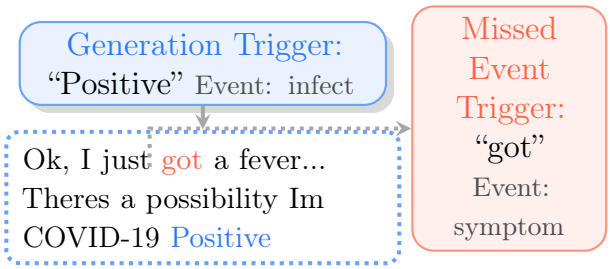

Figure 6.6: Illustration of how inverse generation can produce unannotated event mentions. Blue box = target event mention, red box = unannotated event mention.

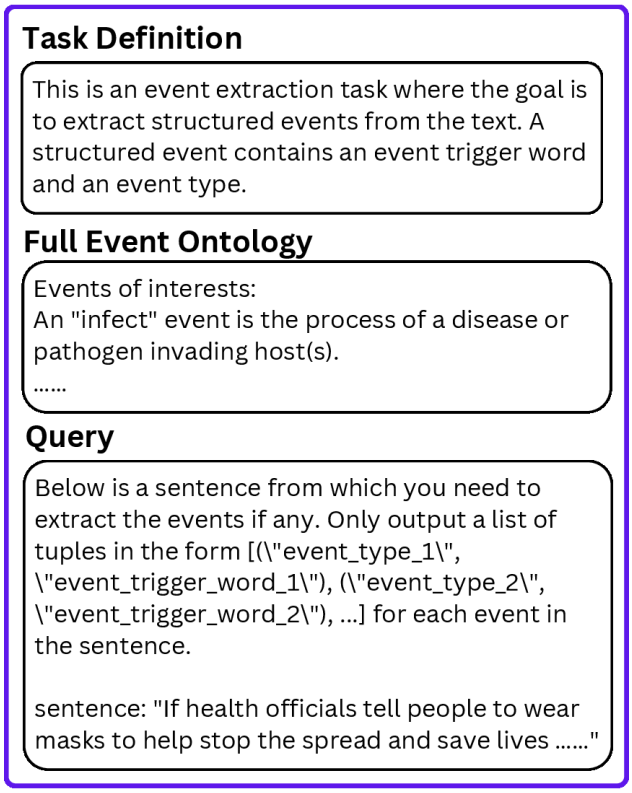

Figure 6.7: Prompt for Refiner.

6.2.3 Refiner

While the Narrator ensures the sampled triggers are mentioned in the passage, it could potentially introduce new unknown events in the passage, leading to under-annotated (X', Y) data instances. We illustrate this in Figure 6.6 where the sampled trigger *positive* for *infect* event is mentioned, but the sentence also mentions the missing *symptom* event triggered by *got*.

To account for such missing annotations, we introduce **Refiner** - tasked to annotate missing events in the generated sentence. Here, we simply prompt the LLM to find all event mentions in the generated sentence (illustrated in Figure 6.7) and append them to the original sampled Y. Since these new refined labels can be noisy, we avoid updating them for the already present events from the Scout, and only add them for newly discovered events. To further improve data quality, we apply an automated rule to remove passages that do not mention the target trigger. Additionally, we standardize trigger annotations by correcting

variations in trigger word forms. Such normalization and conservative filtering further aid in ensuring high data quality. Finally, we apply a greedy sampling algorithm to sample $N = 50$ instances (X', Y) per event type to create our final synthetic dataset D_s.

Downstream Model Training: The final component utilizes the generated synthetic data D_s to train downstream ED models in a supervised manner. The trained ED models are then used to infer on the test set and for eventual evaluation. Since we use small language models, our inference time computation is negligible compared to LLM inference methods.

6.3 Experimental Setup

Datasets: We consider three ED datasets from diverse domains for our experiments: (1) ACE [37], in the news domain, (2) SPEED [164], in the social media domain, and (3) GENIA [95], in the biomedical domain. We simplify GENIA by converting the original document-level annotations to sentence-level annotations. We consider the Arabic and Chinese versions of ACE for cross-lingual experiments. For the few-shot setting, we sample k few-shot examples from the training data.

For our unlabeled data, we consider two sources: (1) **Train** - annotation-free training splits (i.e., only the text) of each dataset and (2) **External** - unlabeled data from other external sources. For this external data source, for ACE, we utilize News Category Dataset [142] comprising Huffpost news articles from 2012-2022. We filter articles corresponding to political, financial, and business articles. For SPEED, we utilize COVIDKB [257], comprising tweets from the Twitter COVID-19 Endpoint. Finally, we utilize GENIA2013 [96]. We provide statistics about these datasets in Table 6.1.

Baseline methods: We consider three LLM-based techniques for low-resource ED as the baselines for our work. (1) Inference [56]: LLMs are used to directly infer on the target test data using their reasoning capability. (2) STAR [134]: The state-of-the-art generation model

Data Source	# Sents	# Event Mentions	Average Length
Test Data			
ACE - test	832	403	22.9
SPEED - test	586	672	28.1
GENIA - test	2,151	1,805	29.7
Unlabeled Train Data			
ACE - train	17,172	-	15.6
SPEED - train	1,601	-	33.5
GENIA - train	6,431	-	30.1
Unlabeled External Data			
HuffPost	43,350	-	17.4
COVIDKB	7,311	-	30.6
GENIA2013	6,542	-	17.4
Multilingual Test Data			
ACE - Arabic	313	198	24.6
ACE - Chinese	486	211	44.2
Unlabeled Multlingual Train Data			
ACE - Arabic	3,218	-	26.1
ACE - Chinese	6,301	-	45.5

Table 6.1: Data Statistics for the various test and unlabeled datasets used for benchmarking our proposed model SNaRe. # = Number of.

for ED utilizing LLMs for trigger and passage generation without using any unlabeled data, (3) Weak Supervision (Weak Sup) [36]: LLMs are utilized to synthesize labels for unlabeled data. For an upper bound reference, we also include a Gold Data generation baseline wherein we sample from the gold training data of each dataset to train the downstream ED model.

Base models: For our base LLMs, we consider three instruction-tuned LLMs of varying sizes, namely Llama3-8B-Instruct (8B model), Llama3-70B-Instruct (70B model) [43], and GPT-3.5 (175B model) [19]. For our downstream ED model, we consider a specialized low-resource model DEGREE [73], a generative model prompted to fill event templates powered by a BART-large pre-trained language model [109].

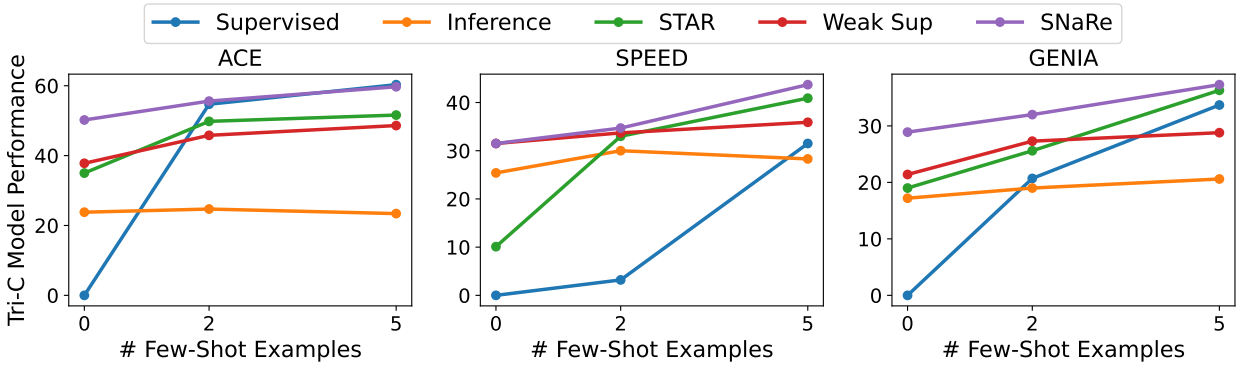

Figure 6.8: Few-shot results comparing SNaRe with other baselines across three datasets using Llama3-8B-Instruct as the base LLM. Except for Inference, all other evaluations are performances of the downstream DEGREE [73] model trained on data generated by each technique. Tri-C: Trigger Classification F1, #: Number of.

Evaluation: Our primary evaluation metric is supervised model performance trained on the synthesized data. We consider two low-resource settings - zero-shot (no labeled data) and few-shot (k datapoints per event type are used). For Inference baseline, the LLM is directly run on the test set to procure model predictions. We report the F1 scores for two metrics [7]: (1) Event Identification (EI) - correct identification of events, and (2) Trigger Classification (TC) - correct identification of trigger-event pairs.

Implementation Details: We follow STAR for the implementation of the baseline models and most hyperparameter settings. For SNaRe's passage generation, we select the top $t = 10$ triggers (except $t = 8$ for GENIA) for passage generation. We generate $N = 50$ datapoints per event type for each generation strategy. All our experimental results are reported over an average of three runs.

6.4 Results and Analysis

Here, we now present the results for our zero-shot, few-shot settings, and cross-lingual experiment below.

Base LLM	Method	Unlabeled Data Source	ACE		SPEED		GENIA		Average	
			EI	TC	EI	TC	EI	TC	EI	TC
Llama3-8B	Inference	-	30.2	23.8	39.8	25.4	21.9	17.2	30.6	22.1
	STAR	-	44.9	35.0	21.0	10.1	25.9	19.0	30.6	21.4
	Weak Sup	train	41.7	37.8	45.6	31.5	26.9	21.4	38.1	30.2
	SNaRe (ours)	train	57.4	50.2	44.6	31.5	**35.2**	**28.9**	45.7	**36.9**
	SNaRe (ours)	external	**57.7**	**52.6**	**47.8**	**32.9**	33.6	24.6	**46.4**	36.7
Llama3-70B	Inference	-	46.9	41.3	46.9	35.6	34.2	28.2	42.7	35.0
	STAR	-	50.0	42.3	18.3	13.8	23.3	16.9	30.5	24.3
	Weak Sup	train	53.2	48.0	**52.8**	**39.6**	36.2	29.1	47.4	38.9
	SNaRe (ours)	train	58.1	53.8	49.9	38.7	38.0	29.7	48.7	40.7
	SNaRe (ours)	external	**59.7**	**55.6**	50.1	39.2	**39.2**	**31.5**	**49.7**	**42.1**
GPT-3.5	Inference	-	33.0	26.2	44.2	32.9	31.2	24.7	36.1	27.9
	STAR	-	45.0	36.6	21.3	14.6	21.8	14.3	29.4	21.8
	Weak Sup	train	49.7	44.6	**50.7**	**37.5**	37.7	30.1	46.1	37.4
	SNaRe (ours)	train	**54.8**	48.3	50.3	36.8	**39.3**	**31.1**	**48.1**	**38.7**
	SNaRe (ours)	external	54.0	**48.5**	50.1	36.1	38.7	29.4	47.6	38.0
(Upper Bound)	Gold Data	-	64.6	61.6	64.0	53.5	51.3	44.0	60.0	53.0

Table 6.2: Zero-shot results comparing downstream performance for data generated by SNaRe with other baselines across three datasets and three base LLMs. Except for Inference, all other evaluations are performances of the downstream DEGREE [73] model trained on data generated by each technique (50 datapoints per event type).

6.4.1 Zero-shot Results

We present the main zero-shot results comparing the baselines across different LLMs and datasets in Table 6.2 and discuss our findings below.

SNaRe performs the best: On average, SNaRe outperforms STAR by 17.3% Eve-I F1 and 16.3% Tri-C F1 points – demonstrating how domain-specific cues from unlabeled data aid in tackling domain drift. Compared to Weak Supervision, SNaRe provides average gains of 3.6% Eve-I F1 and 3.3% Tri-C F1, suggesting that cleaner label quality can help improve model performance.

External data source is effective: Assuming access to the training data as the unlabeled data source can be a strong assumption and bias for SNaRe. To verify the robustness of

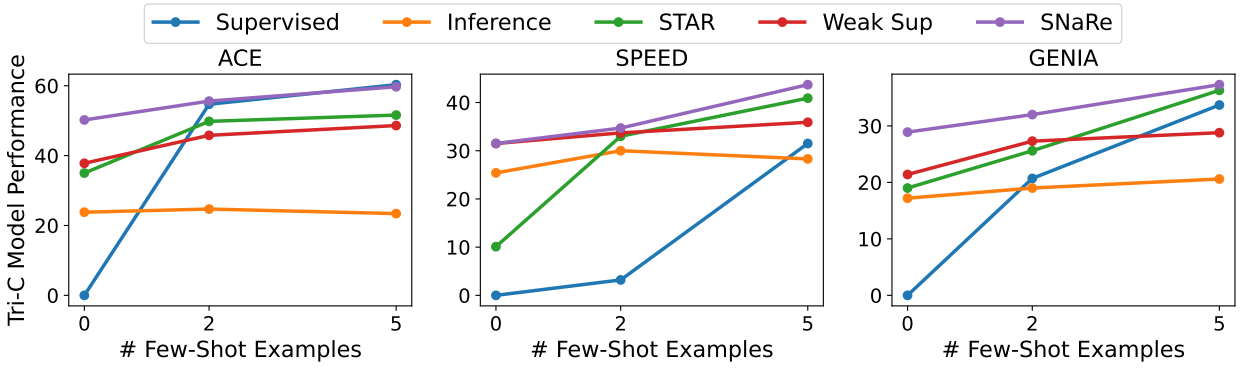

Figure 6.9: Few-shot results comparing SNaRe with other baselines across three datasets using Llama3-8B-Instruct as the base LLM. Except for Inference, all other evaluations are performances of the downstream DEGREE [73] model trained on data generated by each technique. Tri-C: Trigger Classification F1, #: Number of.

our approach, we also evaluate SNaRe with external data sources. Surprisingly, as seen in Table 6.2, SNaRe with external data provides similar gains of 4% Eve-I F1 and 3.4% Tri-C F1 over the best baseline. We posit that the higher volume of external data leads to the extraction of cleaner domain-specific triggers, which eventually aids in better downstream performance.

6.4.2 Few-shot Results

We also study the various methods in the presence of small annotated data as part of our few-shot experiments. Specifically, we study the $k = 2$ and $k = 5$ few-shot settings, where k annotated examples per event type are utilized. We utilize the k-shots as in-context examples in the LLM prompts and append these few-shot examples to the synthesized training data as well. Additionally, we consider another baseline (Supervised) of downstream models trained only on the k-shot examples. We present the Tri-C results for all the datasets for the Llama3-8B model in Figure 6.9. Similar to zero-shot results, we observe that SNaRe consistently outperforms all other baseline models. On average, SNaRe outperforms STAR and Weak Sup by 5.4% Tri-C F1 and 7% Tri-C F1, respectively.

LLM	Method	Arabic		Chinese	
		EI	**TC**	**EI**	**TC**
Llama3-8B	Inference	21.5	13.4	15.0	11.8
	STAR	11.5	10.5	19.7	16.0
	Weak Sup	21.5	16.2	26.3	19.1
	SNaRe (ours)	**40.1**	**33.6**	**35.9**	**31.1**
Llama3-70B	Inference	37.5	27.7	32.0	29.1
	STAR	37.1	30.9	26.0	22.0
	Weak Sup	30.0	20.4	28.1	26.3
	SNaRe (ours)	**47.5**	**44.0**	**40.4**	**33.1**

Table 6.3: Comparing SNaRe with zero-shot inference for other powerful LLMs for the ACE dataset.

Method	ACE	SPEED	GENIA
SNaRe	50.2	31.5	28.9
– Scout	43.2	27.8	28.2
– Narrator	37.8	31.5	21.4
– Refiner	47.4	23.3	22.0

Table 6.4: Ablation study for SNaRe's Scout, Narrator, and Refiner agents measured as Tri-C F1 performance across the three datasets.

6.4.3 Zero-shot Multilingual Results

To highlight the utility of our work, we apply our work across languages, specifically Arabic (ar) and Chinese (zh). We used multilingual ACE data [37] for this experiment and utilized TagPrime [74], powered by XLM-Roberta-large [28], as the downstream ED model. We present our results for Llama3-8B-Instruct and Llama3-70B-Instruct LLMs in Table 6.3. Surprisingly, SNaRe performs the best out-of-the-box, with improvements ranging between 10-20% F1 for Arabic and 4-12% F1 for Chinese, highlighting the broader impact of our work.

6.4.4 Ablation study

Table 6.4 shows the ablation study for Scout and Refiner. For ablating Scout, we replace it by prompting LLM to directly generate triggers. Ablating Narrator is similar to the Weak Sup baseline, while adding additional refiner and dataset statistics to remove noisy datapoints.

LLM + Method	Eve-I	Tri-C
Llama3-70B + Inference	46.9	41.3
Llama3.3-70B + Inference	48.9	43.0
Qwen2.5-72B + Inference	40.9	34.2
GPT4o-mini + Inference	34.5	28.8
GPT4o + Inference	51.4	47.7
QwQ-32B + Inference	49.7	43.5
Deepseek-R1-L3-70B + Inference	41.8	36.6
Llama3-70B + SNaRe (train)	**58.1**	**53.8**
Llama3-70B + SNaRe (external)	**59.7**	**55.6**

Table 6.5: Comparing SNaRe using Llama3-70B with zero-shot inference for other reasoning-based LLMs for the ACE dataset.

Method	ACE	SPEED	GENIA
STAR	9.6%	15.3%	15.1%
Weak Supervision	19.1%	38.4%	44.8%
SNaRe	**23.2%**	**49.4%**	**52.5%**

Table 6.6: Reporting the hit rate of synthesized data triggers relative to gold test triggers for various data generation methods on different domains and datasets.

We observe how all the agents are critical, with average performance reductions of 3.8% F1, 6.6% F1, and 6% F1 upon removing the respective agents of Scout, Narrator, and Refiner.

6.4.5 Comparison with Reasoning LLMs

To further highlight the efficacy of SNaRe, we compare it with zero-shot inference using more recent reasoning-based LLMs for the ACE dataset in Table 6.5. As seen, Llama3-70B with SNaRe significantly outperforms stronger LLMs like GPT4o and thinking-based models like QwQ-32B by 8-10% Eve-I F1 and 8-12% Tri-C F1 scores, respectively. This highlights the strong efficacy of synthetic generation SNaRe over zero-shot inference even when using more powerful LLMs.

Method	Naturalness	Event Relevance	Annotation Quality
STAR	3.1	3.4	3.1
Weak Supervision	**4.2**	-	2.9
SNaRe	3.6	**4.0**	**3.6**

Table 6.7: Human evaluation for sentence naturalness, relevance of event in generated sentence, and the annotation quality for the different data generation methods. 1 = worst, 5 = best.

6.4.6 Analyzing domain drift and label quality

ED models have a strong tendency to over-rely on lexical relations between triggers and events [212]. Thus, we compare the synthetic data triggers with the gold test triggers as a raw study of the domain drift of triggers in the synthesized data. Specifically, we extract triggers per event type in the synthetic datasets and measure the hit rate/overlap of the synthesized triggers with the gold set of triggers, as reported in Table 6.6. STAR's low hit rate indicates the poor overlap with the gold triggers, which is a primary reason for its domain drift. Furthermore, the consistently stronger coverage of SNaRe explains its lower domain drift.

To further study the quality and relevance of the label, we performed a human evaluation. Specifically, a human expert in ED is tasked with scoring generations (between 1-5) on the naturalness of the sentence specific to the target domain, the relevance of the event to the semantic actions described in the generated sentence, and the annotation quality evaluating the selection of triggers. We provide the averaged scores across the three datasets for 90 samples in Table 6.7.[1] Weak Supervision has high sentence quality but poor label annotations; STAR suffers from poor event relevance, indicating domain drift. Overall, SNaRe performs the best with high annotation quality and event relevance.

[1]Since Weak Supervision annotates the unlabeled target domain sentences, event relevance is analogous to annotation quality, and we do not explicitly evaluate it.

Method	ACE		SPEED		GENIA	
	EI	**TC**	**EI**	**TC**	**EI**	**TC**
SNaRe	**57.4**	50.2	44.6	31.5	35.2	28.9
+ SFT LLM	55.2	**51.7**	**46.9**	**35.8**	**36.7**	**29.1**

Table 6.8: Measuring performance improvement by fine-tuning an LLM on unlabeled train data for SNaRe.

Dataset	Event	Method	Trigger	Sentence
ACE	Attack	STAR	raid	As the rebels embarked on a daring trek across the desert, they launched a surprise **raid** on the heavily guarded fortress, catching the enemy off guard.
		SNaRe	shooting	As the rival businessman signed the contract, a sudden **shooting** erupted outside, causing chaos in the midst of the transaction.
SPEED	Death	STAR	asphyxiation	The hiker's life was tragically cut short as **asphyxiation** occurred after she became stuck in the narrow cave crevice.
		SNaRe	killed	The patient's feverish state was triggered when they tested positive for the virus, which ultimately led to their being **killed** by the rapidly spreading infection.
GENIA	Binding	STAR	merge	The regulatory protein's ability to activate a specific region of the DNA triggers the **merge** of two proteins, leading to the modification of gene expression.
		SNaRe	bound	During the phosphorylation of the enzyme, it **bound** to the DNA sequence, initiating the transcription process.

Table 6.9: Qualitative examples demonstrating STAR and SNaRe's trigger and sentence generation quality.

6.4.7 Domain-adapted LLM Fine-tuning

We fine-tune the base LLM for Narrator on the unlabeled target domain train data D'_T to better align the generated passages. Naturally, this can be applied only for smaller LLMs owing to fine-tuning costs. We present the results of fine-tuning Llama3-8B-Instruct on the unlabeled train data in Table 6.8. On average, we observe that target data fine-tuning additionally improves SNaRe by 0.5-2% F1. Qualitative studies indicate that the generated passages are distributionally closer to the target domain, further reducing domain drift.

6.4.8 Qualitative analysis of generated data

We provide qualitative evidence for SNaRe's reduction in domain drift by Scout's domain-specific triggers in Table 6.9. We compare with STAR, which uses LLM's internal knowledge to generate the triggers. Lack of domain grounding often results in STAR's triggers and sentences being misaligned (e.g. *asphyxiation* for *death* event related to pandemics) relative to the target domain. In contrast, SNaRe's triggers are better aligned to the target domain corpus, resulting in better quality data and reduced domain drift.

6.5 Conclusion

We introduce SNaRe, a domain-aware synthetic data generation approach composed of Scout, Narrator, and Refiner. Utilizing Scout's domain-specific triggers for synthesizing sentences, along with Narrator's conditional generation and Refiner's annotations, helps reduce domain drift and label noise. Experiments on three diverse datasets in zero-shot, few-shot, and multilingual settings demonstrate the efficacy of SNaRe, establishing SNaRe as a strong data generation framework.

Chapter 7

SPEED: A Multilingual Event Extraction Framework for Epidemic Preparedness using Social Media

Early warnings and effective control measures are among the most important tools for policy-makers to be prepared against the threat of any epidemic [27]. World Health Organization (WHO) reports suggest that 65% of the first reports about infectious diseases and outbreaks originate from informal sources and the internet [71]. Social media is an important information source here, as it is more timely than other alternatives like news and public health [103], more publicly accessible than clinical notes [132], and possesses a huge volume of content.[1] This underscores the need for an automated system monitoring social media to provide early and effective epidemic prediction, tying back to the application motivation of this thesis.

To this end, in this chapter, we introduce **SPEED** and **SPEED++** (**S**ocial **P**latform based **E**pidemic **E**vent **D**etection), the first multilingual EE framework designed for epidemic preparedness using social media. Compared to existing simple epidemiological keyword and sentence-classification approaches [108, 132], EE requires a deeper semantic understanding.

[1] A daily average of 20 million tweets were posted about COVID-19 from May 15 – May 31, 2020.

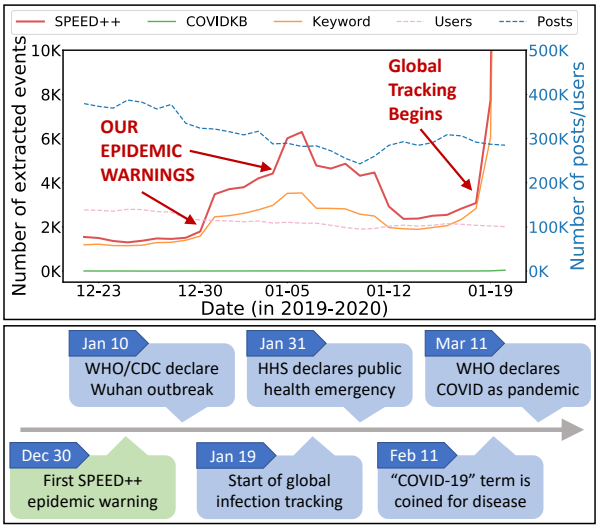

Figure 7.1: Zero-shot multilingual epidemic prediction in Chinese for COVID-19 pandemic. (Top) Number of epidemic events extracted in Dec-Jan 2020. Arrows indicate SPEED epidemic warnings. (Bottom) SPEED warning with respect to the general timeline of major moments of the COVID-19 pandemic.

This enhanced understanding aids in more effective disease-agnostic extraction of epidemic events from social media. By reporting sharp increases in epidemic-related events, we can provide early epidemic warnings, as shown in Figure 7.1, where we detect COVID-19 epidemic warnings zero-shot in Chinese - highlighting the applicability of ED for epidemic prediction.

Existing EE datasets are unsuitable for establishing a framework to extract epidemic-related events from social media, as they focus on general-purpose events in news and Wikipedia domains, while other epidemiological works are disease-specific and too fine-grained, as we discuss in § 7.1. Instead, we construct our own epidemic EE ontology and dataset for social media. Our created ontology comprises seven event types - *infect*, *spread*, *symptom*, *prevent*, *cure*, *control*, *death* - and 20 argument roles (e.g. *symptoms*, *control measures*, *infected*, etc.) chosen based on their relevance for epidemics, frequency in social media, and their applicability to various diseases. We further validate our ontology through clinical sources and public health experts.

For the dataset, we chose Twitter as the social media platform and focused on the COVID-19 pandemic. Apart from traditional English, we also annotate on three other languages

- Spanish, Hindi, and Japanese - to benchmark multilingual EE models. Leveraging the enriched ontology and expert annotations, we develop our SPEED/SPEED++[2] dataset comprising 5.1K tweets and 4.6K event mentions across four different diseases (COVID-19, Monkeypox, Zika, and Dengue) in four languages. Using our SPEED/SPEED++ dataset, we develop English-only and zero-shot cross-lingual models. For English-only models, we train existing ED frameworks [42, 73]; while for zero-shot cross-lingual models, we empower TagPrime [74] with multilingual pre-training and augmented training using pseudo-generated multilingual data from CLaP [162]. These models are trained realistically on limited English COVID-specific data and expected to provide predictions cross-disease, cross-language, in a zero-shot manner - a more realistic evaluation. Benchmarking on SPEED reveals how our trained models outperform various baselines by an average of 10-16% F1 points for unseen diseases across four different languages.

To demonstrate the utility of our multilingual EE framework, we apply it to two epidemic-related applications. First, we utilize the framework's cross-disease extraction for early epidemic prediction by aggregating epidemic events across time. Comparing our extracted events with the actual reported cases, we show that our English-only cross-disease framework can provide warnings up to 4-9 weeks earlier than the WHO declaration for the Monkeypox epidemic (Figure 7.2). Applying our cross-lingual framework for COVID-19 to Chinese Weibo posts, we successfully detected early epidemic warnings by Dec 30, 2019 (Figure 7.1) - three weeks before the global infection tracking even began. By incorporating tweet locations, we construct a global epidemic severity meter capable of providing epidemic warnings in 65 languages spanning 117 countries. Such early warnings, aided by timely action, can potentially lead to a 2-4x reduction in the number of infections and deaths [93] - signifying the role of our framework for global epidemic prediction.

As another application, we repurpose our framework as an information aggregation system for community discussions about epidemics such as *symptoms, cure measures*, etc. Leveraging

[2]Note that this dataset was created in two phases. The first phase SPEED focused on English ED and the second phase SPEED++ focused on Multilingual EAE annotations.

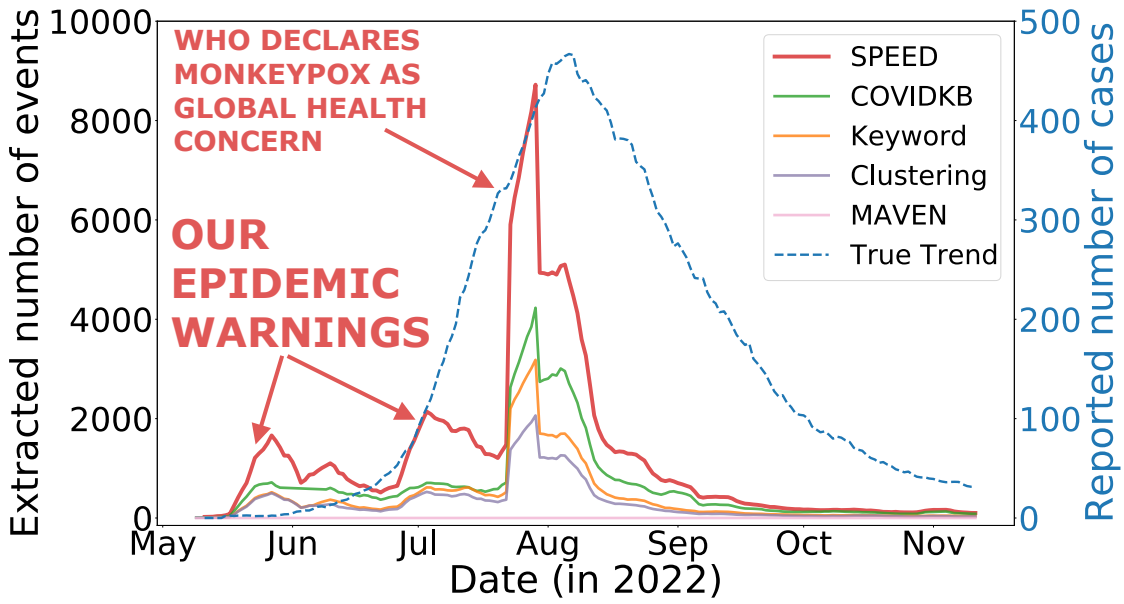

Figure 7.2: Number of reported Monkeypox cases and extracted events by our trained ED model from May 11 to Nov 11, 2022. Arrows indicate how our system could provide early epidemic warnings about 4-9 weeks before the WHO declared Monkeypox as a concern. MAVEN = Data Transfer model trained on MAVEN. Keyword = epidemiological keyword baseline.

the EAE capability of our framework, we meticulously extract these event-specific details from millions of tweets across diseases and languages. Similar arguments are then agglomeratively clustered to generate an aggregated ranked bulletin. We demonstrate that this bulletin can aid misinformation detection (e.g., *cow urine* as a cure for *COVID-19*) and public attention shift monitoring (e.g., *rashes* as symptoms for *Monkeypox*). Such an automated disease-agnostic multilingual aggregation system can significantly alleviate human effort while providing insights into public epidemic opinions.

Overall, we discuss two works in this chapter, pioneering the utilization of Event Extraction in social media for Epidemic Preparedness. In turn, we make the following wholistic contributions: (1) we create a disease-agnostic social-media-tailored ontology and the first multilingual Event Extraction dataset for epidemic prediction SPEED/SPEED++ encompassing four diseases and four languages, (2) we extensive benchmark existing methods to demonstrate their inadequacy and the substantial improvements achieved by our pro-

posed models trained on SPEED, (3) we demonstrate the robust utility of our framework through two epidemic-centric applications, facilitating multilingual epidemic prediction and the aggregation of epidemic information.

We organize our work by discussing the background and related works in Section 7.1. Next, we discuss our process to create the ontology, annotate the data, along with some data analysis in Section 7.2. Then, we discuss our first major results for zero-shot Event Extraction for epidemiology in Section 7.3. Post this, we provide two applications of our work in Section 7.4 and Section 7.5. This chapter has been published as two works - Parekh et al. [164], and Parekh et al. [163].

7.1 Background and Related Work

Here, we provide a brief background of the shortcomings of existing Event Extraction (EE) datasets. We also discuss relevant epidemiological ontologies, datasets, and developed information extraction systems.

Event Extraction Datasets Earliest works for this task can be dated back to MUC [204, 61] and the more standard ACE [37]. Over the years, ACE was extended to various datasets like ERE [196] and TAC KBP [47, 48]. Recent progress has been the creation of massive datasets and huge event ontologies with datasets like MAVEN [220], RAMS [45], WikiEvents [121], DocEE [212], GENEVA [161] and GLEN [122]. While most of these datasets are only in English, some datasets like ACE [37], ERE [196], and MEE [171] provide EE data in ten languages for general-purpose events in the news and Wikipedia domains. Overall, these ontologies and datasets cater to general-purpose events, are primarily English, and are not suitable for multilingual epidemiological EE.

Epidemiological Ontologies Early epidemiological works [126, 176] defined highly rich taxonomies for describing technical concepts used by biomedical experts. Further develop-

Dataset	Source	Sentence Level	Trigger Present	Social Events	Personal Events	Social Media Granular	Disease agnostic
COVIDKB [257]	Twitter	✔	✗	✗	✔	✔	✗
CACT [132]	Clinical	✗	✔	✗	∼	✔	✗
ExcavatorCovid [140]	News	✗	✔	✔	✔	✗	✗
BioCaster [27]	News	✗	✗	✔	✔	✗	✔
DANIEL [108]	News	✗	∼	✗	∼	✔	✔
SPEED (Ours)	Twitter	✔	✔	✔	✔	✔	✔

Table 7.1: Objective comparison of various epidemiological datasets COVIDKB [257], CACT [132], ExcavatorCovid [140], BioCaster [27], and DANIEL [108] with our dataset SPEED. We objectify the source of data (Data Source), the level of annotation granularity (Sentence Level), the presence of trigger information (Trigger Present), the presence of social and personal events (Social Events and Personal Events), the suitability of ontology for social media (Social Media Granular), and whether the dataset is disease-agnostic (Disease-agnostic). ✔ indicates present, ✗ indicates absence, and ∼ indicates partial presence.

ments led to the creation of SNOMED CT [199] and PHSkb [40] that define a list of reportable events used for communication between public health experts. BioCaster [27] and PULS [41] extended ontologies for the news domain. Recent works of NCBI [38], IDO [13], and DO [189] focus on comprehensively organizing human diseases. In light of the recent COVID-19 pandemic, CIDO [69] defines a technical taxonomy for coronavirus, while ExcavatorCovid [140] automatically extracts COVID-19 events and relations between them. Most of these ontologies are too fine-grained or limited to specific events, and can't be directly used for ED from social media, as also shown in Table 7.1.

Epidemiological Information Extraction Early works utilized search-engine queries and click-through rates for predicting influenza trends [49, 60]. Information extraction from Twitter has also been quite successful for predicting influenza trends [194, 103, 169]. Over the years, various biomedical monitoring systems have been developed, like BioCaster [27, 139], HeathMap [55], DANIEL [108], EpiCore [152]. Extensions to support multilingual systems have also been explored [108, 144, 181]. For the COVID-19 pandemic, several frameworks like CACT [132] and COVIDKB [257] were developed for extracting symptoms and infection statistics, respectively. While most of these works are English-focused, some other works

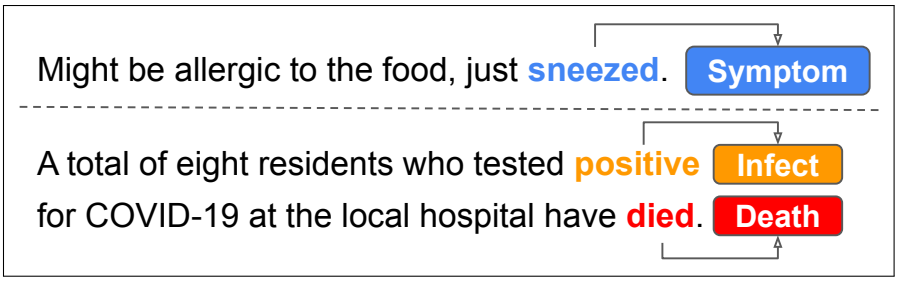

Figure 7.3: Illustration for the task of Event Detection for detecting epidemic-based events. Event mentions: Event *symptom* and trigger *sneezed* (1st sentence), Event *infect* and trigger *positive* (2nd sentence), Event *death* and trigger *died* (2nd sentence).

[108, 144, 181] support multilingual systems by using keyword-based and simple classification methods. Overall, most of these systems are English-centric, disease-specific, focus on news and clinical domains, and use keyword/rule-based or simple BERT-based models, as shown in Table 7.1. In our work, utilizing disease-agnostic annotations and powerful multilingual models, we develop models that can detect events for any disease mentioned in any language.

7.1.1 From Event Detection to Epidemic Prediction

Given a social media post, Event Detection (ED) [204, 37] extracts and classifies significant events of interest. By designing disease-agnostic epidemic-based events, we aim to train ED models to extract epidemic events from social media posts for any possible disease, as illustrated in Figure 7.3 for three epidemic events *symptom*, *infect*, and *death*. By detecting abnormal influx in the trends of extracted epidemic events from social media, we can thus provide early epidemic warnings for any possible disease, as we show for Monkeypox in Figure 7.2. Existing epidemiological approaches [108, 132] are simple keyword or sentence classification-based and generally less accurate. Other works like COVIDKB [257] and ExcavatorCovid [140] are disease-specific and utilize events for building knowledge bases, as shown in Table 7.1. To the best of our knowledge, we are the first ones to leverage event detection to extract epidemic events from social media and provide early warnings for any possible disease.

7.2 Ontology Creation and Data Collection

We choose social media as our document source as it provides faster and more timely worldly information than news and public health [103] and is more publicly accessible than clinical notes [132]. Owing to its public access and huge content volume, we consider **Twitter**[3] as the social media platform and consider the recent **COVID-19 pandemic** as the primary disease. Since existing epidemiological ontologies and standard ED datasets can't be utilized to exploit ED for epidemic prediction (§ 7.1 and Table 7.1), we create our own event ontology and dataset SPEED for detecting disease-agnostic epidemics from social media. Furthermore, we extend to three other languages apart from English - Spanish, Hindi, and Japanese - to enhance the multilingual capability of our framework. We construct our ontology/dataset in two stages - in the first stage, we create an English-centric ED ontology, and in the second stage, we augment it with EAE components and multilingual data. Figure 7.4 provides a brief overview of our first stage, while Figure 7.5 provides a high-level overview of our second stage.

7.2.1 Ontology Creation

First, we create our event ontology for the task and describe the steps as follows:

1. **Curation of event types**: We scan through existing medical ontologies like BCEO [27], IDO [13], and the ExcavatorCovid [140] and curate a large list of event types for infectious and epidemic-related diseases.

2. **Merge event types across ontologies**: Since these existing ontologies may have repetitive event types, we perform a merging step. Specifically, two human experts manually examine and merge event types that are exactly similar in our curated list of event types (e.g. *Outbreak* event type).

[3] https://www.twitter.com/

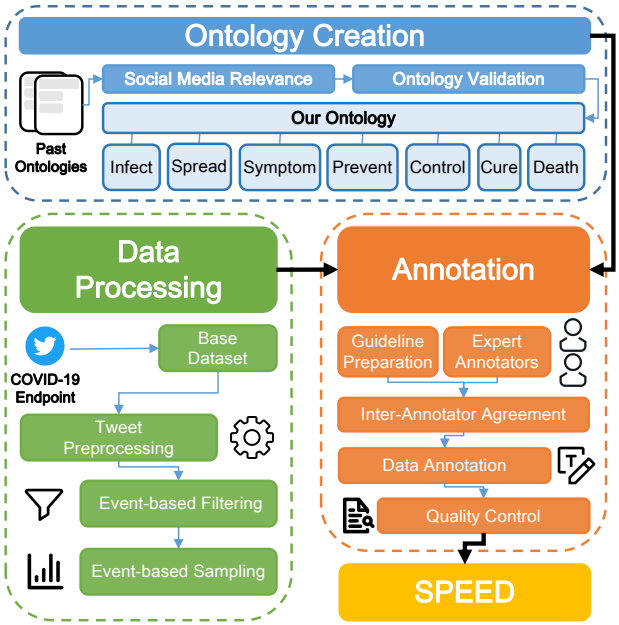

Figure 7.4: Overview of the first stage of our SPEED dataset creation process with three major steps: Ontology Creation, Data Processing, and Data Annotation.

3. **Filter out disease-specific event types**: Some event types in our curated list are specific to certain diseases. We identify and filter out such event types (e.g. Mask Wearing for COVID-19 which may not be observed for other diseases). We utilize opinions from public health experts to aid this step ensuring our event types are disease-agnostic.

4. **Definition Correction**: Utilizing aid from public health experts, we add and refine definitions for the curated set of event types and ensure they are disease-agnostic.

5. **Organization**: Following ExcavatorCovid [140], we organize our curated list of event types into three larger categories: social (events involving larger populations), personal (individual-oriented events), and medical (medically focused events) types.

Our complete initial event ontology comprises 18 event types along with their event definitions organized into three abstract categories as shown in Table 7.2.

To further enhance our ontology with argument roles, we initially drafted event-specific roles through a crowdsourced survey with 100 participants. Through manual inspection, we

Event name	**Definition**	**Action for Final Ontology**
	Social Scale Events	
Prefigure	The signal that precedes the occurrence of a potential epidemic.	Discarded
Outbreak	The process of disease spreading among a certain amount of the population at a massive scale.	Merged into *Spread*
Spread	The process of disease spreading among a certain amount of the population but at a local scale.	Final Event
Control	Collective efforts trying to impede the spread of a epidemic.	Final Event
Promote	The relationship of a disease driver leading to the breakout of a disease.	Discarded
	Personal Scale Events	
Prevent	Individuals trying to prevent the infection of disease.	Final Event
Infect	The process of a disease/pathogen invading host(s).	Final Event
Symptom	Individuals displaying physiological features indicating the abnormality of organisms.	Final Event
Treatment	The process that a patient is going through with the aim of recovering from symptoms.	Merged into *Cure*
Cure	Stopping infection and relieving individuals from infections/symptoms.	Final Event
Immunize	The process by which an organism gains immunization against an infectious agent.	Merged into *Prevent*
Death	End of life of individuals due to infectious disease.	Final Event
	Medical Scale Events	
Cause	The causal relationship of a pathogen and a disease.	Discarded
Variant	An alternation of a disease with genetic code-carrying mutations.	Discarded
Intrude	The process of an infectious agent intruding on its host.	Merged into *Infect*
Respond	The process of a host responding to an infection.	Discarded
Regulate	The process of suppressing and slowing down the infection of a virus.	Merged into *Cure*
Transmission route	The process of a pathogen entering another host from a source.	Discarded

Table 7.2: Complete initial epidemic event ontology comprising 18 event types organized into 3 higher-level abstract categories. We also present details about the event definitions and the action taken for each event type in the final ontology.

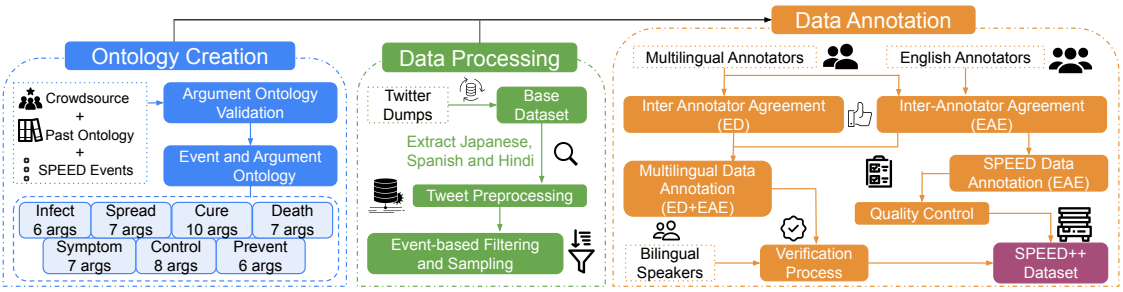

Figure 7.5: Overview of the second stage of our SPEED++ dataset creation process with three major steps: Ontology Creation, Data Processing, and Data Annotation.

extract the frequently mentioned roles in the responses. These are augmented with more typical roles like *Time* and *Place*. We further expand the ontology with epidemic-specific roles (e.g. *Effectiveness of a cure*, *Duration of a symptom*) from a fine-grained COVID ontology ExcavatorCovid [140]. Finally, the roles are renamed to reflect corresponding events (e.g., the person who gets infected in the *infect* event is named *Infected*).

7.2.2 Social Media based Filtering and Validation

Our initial ontology (Table 7.2) was constructed using previous ontologies and human knowledge. But the relevance of each event type for social media (specifically Twitter) remains unknown. To evaluate this relevance, we first associate each event type with event-specific keywords. Then we utilize frequency and specificity as two guiding heuristics for further filtering/merging of event types in our curated ontology. We utilize the base Twitter dataset for conducting this analysis and describe each of these steps in more detail here:

1. **Keyword Association:** To objectively conduct this analysis, we associate each event type with a set of keywords. This association involves two simple steps:

   (a) *Human expert curation*: For each event type, a human expert curates 2-3 simple yet specific keywords for each event based on commonsense knowledge.

   (b) *Thesaurus-based expansion*: For each human-expert curated list, we utilize an

external resource - Thesaurus[4] to further find event-relevant keywords.

2. **Frequency-based filtering:** Using frequency, we aim to filter out event types that are less mentioned in social media. To approximately estimate the frequency of each event type in social media, we count the number of social media posts containing any of the curated keywords for each event type. We observe that most events under the medical abstraction occur much lesser than others. Since such low-frequency events (e.g. *Variant*, *Cause*, *Prefigure*, etc.) are less likely to be mentioned in a smaller sample of data, we discard or merge such events for our final ontology. We repeat this step similarly for event-specific argument roles.

3. **Specificity-based filtering:** Specificity ensures that each event type is uniquely identifiable with a good confidence and mainly aims to reduce ambiguity and make the event types more distinct. To estimate specificity, for each curated keyword of an event type, we randomly sample a small number of non-duplicate social media posts and annotate them with human experts. For example, the *Control* event comprises high specificity keywords such as *quarantine*, *protocol*, *guidelines*. On the other hand, the event *Prefigure* doesn't have any high specificity keywords, but only medium specificity keywords such as *foreshadow* and low specificity keywords such as *foretell*. Since medium and low specificity keywords are more likely to give false positives, we filter/merge event types that have a high number of low-confidence keywords (e.g. *Intrude*, *Promote*).

Thus, with the above filtering and merging, we shrink our ontology from 18 event types to seven event types that are distinguishable, frequent, and have a low false-positive rate. We provide details about the action taken for each event type with respect to the final ontology in Table 7.2.

Ontology Validation and Coverage Elemental medical soundness is ensured for our ontology since it is derived from established epidemiological ontologies. To further certify this

[4] https://www.thesaurus.com/

Event Type	Argument Roles
Infect	infected, disease, place, time, value, information-source
Spread	population, disease, place, time, value, information-source, trend
Symptom	person, symptom, disease, place, time, duration, information-source
Prevent	agent, disease, means, information-source, target, effectiveness
Control	authority, disease, means, place, time, information-source, subject, effectiveness
Cure	cured, disease, means, place, time, value, facility, information-source, effectiveness, duration
Death	dead, disease, place, time, value, information-source, trend

Table 7.3: Final Event Ontology for SPEED++ comprising 7 event types and 20 argument roles.

soundness, two public health experts validate the sufficiency and comprehensiveness of our ontology and event definitions. To verify if our ontology is characteristic of any disease, we assess our ontology coverage for four diverse diseases by estimating the percentage of event occurrence in disease-related tweets. Notably, we observe a high coverage: 50% for COVID-19, 44% for Monkeypox, 70% for Dengue, and 73% for Zika, confirming robust disease coverage of our ontology. The final ontology of event types and roles is presented in Table 7.3.

7.2.3 Data Processing

To access a wide range of tweets related to COVID-19, we utilized the Twitter COVID-19 Endpoint released in April 2020. We used a randomized selection of **331 million tweets** between May 15 - May 31 2020, as our base dataset. We utilize dumps from Dias [35] as the Zika+Dengue base dataset and Thakur [208] for Monkeypox. For preprocessing tweets, we follow Pota et al. [170]: (1) we anonymize personal information like phone numbers, emails, and handles, (2) we normalize any retweets and URLs, (3) we remove emojis and split hashtags, (4) we filter out tweets only in English.

Event-based Filtering Most tweets in our base dataset expressed subjective sentiments, while only 3% comprised mentions aligned with our event ontology.[5] To reduce annotation costs, we further filter these tweets using a simple *sentence embedding* similarity technique. Specifically, each event type is linked to a seed repository of 5-10 diverse tweets. Query tweets are filtered based on their sentence-level similarity [177] with this event-based seed repository.[6] This step filters about 95% tweets from our base dataset, leading to a 20x reduction in the annotation cost.

Event-based Sampling Random sampling of tweets would yield an uneven and COVID-biased distribution of event types for our dataset. We instead perform a uniform sampling - wherein we over-sample tweets linked to less frequent types (e.g. *prevent*) and under-sample the more frequent ones (e.g. *death*). Such a uniform sampling has proven to ensure model robustness [161] - as also validated by our experiments - and in turn, would make SPEED generalizable to a wider range of diseases. In total, we sample 1,975 tweets which are utilized for ED annotation.

7.2.4 English ED Data Annotation

For English-only ED annotation (first stage), annotators are tasked with identifying whether a given tweet mentions any event outlined in our ontology. If an event is present, annotators are required to identify the specific event trigger. We design our annotation guidelines following the standard ACE dataset [37] and amend them through several rounds of preliminary annotations to ensure annotator consistency.

Annotator Details To ensure high annotation quality and consistency, we chose six experts instead of crowdsourced workers. These experts are computer science students studying NLP and well-versed in ED. They were further trained through multiple rounds of annotations

[5]Based on a keyword-based study conducted on 1,000 tweets
[6]We use a filtering threshold of 0.9.

and feedback.

Inter-annotator agreement (IAA) We used Fleiss' Kappa [54] for measuring IAA and conducted two phases of IAA studies:

1. *Guideline Improvement:* Three annotators participated in three annotation rounds to improve the guidelines through collaborative discussion of disagreements. IAA score rose from 0.44 in the first round to 0.59 (70 samples) in the final round.

2. *Agreement Improvement:* All annotators participated in three rounds of annotations to boost consistency. IAA score improved from 0.56 in the first round to a strong 0.65 (50 samples) in the final round.

Quality Control Apart from extensive IAA studies, we deploy two mechanisms to ensure the high annotation quality:

1. *Multi-Annotation:* Each tweet is annotated by two annotators, and disagreements are resolved by a third annotator.

2. *Flagging:* Annotators can "flag" ambiguous annotations, which are then resolved and annotated by a third annotator through collective discussion.

Both these mechanisms, along with a good IAA score, ensure the high quality of SPEED's ED annotations.

7.2.5 Multilingual EAE Data Annotation

We conduct two sets of annotations to create our multilingual EE dataset: (1) EAE annotations for existing SPEED English ED data and (2) ED+EAE annotations for data in Japanese, Hindi, and Spanish. For ED, annotators were tasked to identify the presence of any events in a given tweet. For EAE, annotators were further asked to identify and extract event-specific roles that were also mentioned in the tweet.

Annotators and Agreement To maintain high annotation quality, our annotators were selected to be a pool of seven experts who were computer science NLP students trained through multiple annotation rounds. Of these seven, we had three annotators who were bilingual speakers of English and Japanese/Hindi/Spanish, respectively. These three annotators handled the multilingual ED and EAE annotations. The remaining four annotators, along with the bilingual English-Hindi annotator, focused on English EAE annotations.

To ensure good annotation agreement, we conduct two agreement studies among the annotators: (1) ED annotations for multilingual annotators and (2) EAE annotations for all annotators. Both these studies were conducted using English data (even for multilingual annotations) to ensure that agreement could be measured in a fair manner. Agreement scores were measured using Fleiss' Kappa [54]. For ED agreement, two rounds of study for the 3 multilingual annotators yielded a super-strong agreement score of 0.75 (30 samples). For EAE, the agreement score for the 7 annotators after two annotation rounds was a decent 0.6 (25 samples).

Annotation Verification To mitigate single annotator bias, each datapoint in the English data is annotated by two annotators, with a third annotator resolving inconsistencies. Owing to the scarcity of multilingual annotators, we hire three additional bilingual speakers to verify the multilingual annotations. These verification annotators were selected through a thorough qualification test to ensure high verification quality. They were requested to judge if the current annotations were reasonable. If the original annotation was deemed incorrect, they were asked to provide feedback to correct the annotations. This feedback was finally utilized by the original multilingual annotators to rectify the annotation.

7.2.6 Data Analysis

Comparison with other datasets SPEED/SPEED++ comprises 5,106 tweets with 4,674 event mentions and 13,815 argument mentions across four diseases and four languages. We

Dataset	# Langs	# ET	# Arg Roles	# Sent	# EM	Avg. EM	# Args	Avg. Args	Domain
Genia2013	1	13	7	664	6,001	429	5,660	809	Biomedical
MLEE	1	29	14	286	6,575	227	5,958	426	Biomedical
ACE	3	33	22	29,483	5,055	153	15,328	697	News
ERE	2	38	21	17,108	7,284	192	15,584	742	News
MEE	8	16	23	31,226	50,011	3126	38,748	1685	Wikipedia
SPEED++	4	7	20	5,107	4,677	668	13,827	691	Social Media

Table 7.4: Data statistics for SPEED/SPEED++ dataset and comparison with other standard EE datasets. Langs = languages, # = number of, ET = Event Types, Avg. = average, Sent = sentences, EM = event mentions, Args = arguments.

Lang	# Sent	Avg. Length	# EM	# Args
en	2,560	32.5	2,887	8,423
es	1,012	32.4	614	1,485
hi	716	30.0	627	2,344
ja	819	89.2*	549	1,575

Table 7.5: Data statistics for SPEED/SPEED++ split by language. # = number of, Avg = average, Lang = language, Sent = sentences, Args = arguments, *character-level.

present the main statistics along with comparisons with other prominent EE datasets like ACE [37], ERE [196], Genia2013 [117], MEE [171], and MLEE [173] in Table 7.4. We note that SPEED++ is one of the few multilingual EE datasets, notably the first in the social media domain. Overall, SPEED++ is comparable in various event and argument-related statistics with the previous standard EE datasets.

Multilingual Statistics We provide a deeper split of data statistics per language in Table 7.5. Owing to cheaper annotations, English has many more annotated sentences compared to other languages. This is also a design choice, as we will solely utilize English data for training zero-shot multilingual models. In terms of event and argument densities (i.e. # EM / # Sent and # Args / # Sent), we notice a broader variation across languages, with English and Hindi being denser. The average lengths (in terms of the number of words) are similar across the languages.

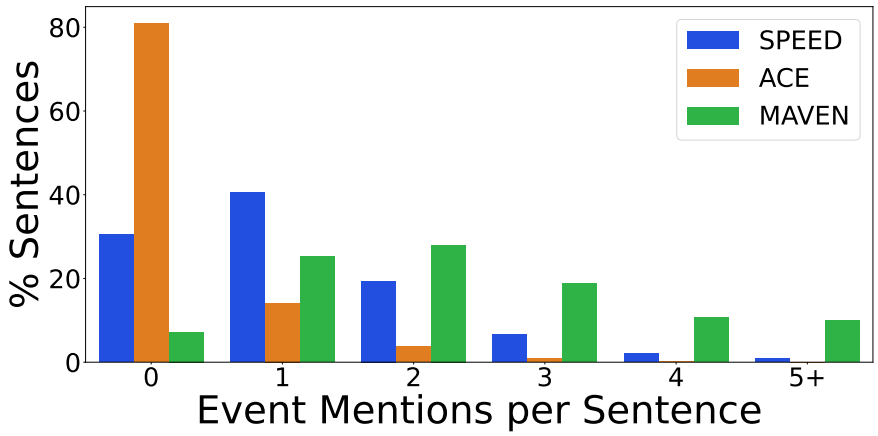

Figure 7.6: Distribution of number event mentions per sentence for SPEED, ACE, and MAVEN. Here % indicates percentage.

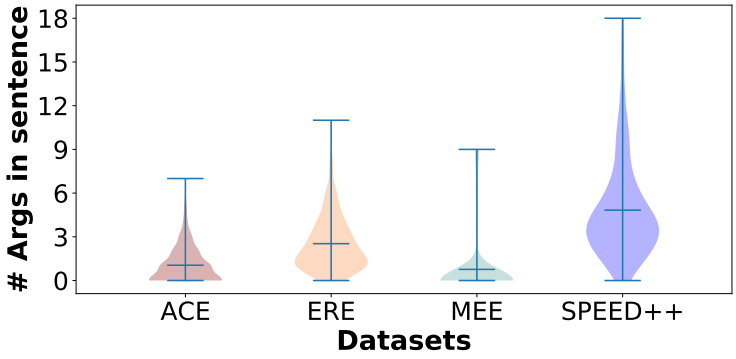

Figure 7.7: Distribution of the number of arguments (# Args) per sentence for SPEED/SPEED++ relative to other datasets ACE, ERE, and MEE.

Comparable Datasize Since we only focus on 7 event types, SPEED has a relatively smaller number of sentences and event mentions. However, SPEED has a high 668 average mentions per event type (column 7 in Table 7.4), more than most other standard datasets. We compare the distribution of event mentions per sentence with other ED datasets like ACE and MAVEN in Figure 7.6. We observe that the event density of our dataset is less than MAVEN but better than ACE. This shows that SPEED is a fairly dense and reasonably sized ED dataset.

Argument Role Study We deep-dive to study the density in terms of arguments per sentence by comparing with standard EE datasets ACE and ERE, and a multilingual EE

	Disease	Language	# Sent	# EM
Train	COVID	English	1,601	1,746
Dev	COVID	English	374	471
Test	COVID	Spanish	534	365
		Hindi	416	412
		Japanese	542	395
	Monkeypox	English	286	398
	Zika + Dengue	English	299	272
		Spanish	478	249
		Hindi	300	215
		Japanese	277	154

Table 7.6: Data split for epidemic event extraction using SPEED/SPEED++. # = number of, Sent = sentences, EM = event mentions.

dataset MEE in Figure 7.7. Noticeably, SPEED/SPEED++ is more dense (has a higher mean argument per sentence value) and has a broader distribution with sentences up to 18 arguments as well. Furthermore, following GENEVA [161], we add 4 non-entity roles, which make up 20% of the total arguments. Such non-entity arguments are not present in any other multilingual datasets. Overall, the high and broad argument density and the existence of non-entity arguments render SPEED/SPEED++ to be a more challenging EE dataset.

7.3 Zero-shot Event Extraction

To validate the effectiveness of EE for epidemic events, we benchmark various EE models using SPEED/SPEED++. We focus on the setting of zero-shot cross-disease evaluation to realistically benchmark the practical utility of our framework. Specifically, we focus on training the models on COVID-19 and evaluating them on Zika/Dengue/Monkeypox. These diseases are fairly distinct, as Monkeypox causes rashes and is rarely fatal, Zika causes birth defects, and Dengue causes high fever and can be fatal. Given the infeasibility of procuring quality data in all languages for all diseases, we benchmark in a zero-shot cross-lingual cross-disease fashion, i.e., we train models only on English COVID data and evaluate on the rest. We provide the data split for our benchmarking in Table 7.6. For evaluation, we report

F1-score for event classification (**EC**) and argument classification (**AC**) [7], measuring the classification of events and arguments, respectively.

7.3.1 EE Models

Most EE models solely focus on English and can not be directly utilized in the cross-lingual setting. To this end, we adopt the following models using multilingual pre-trained models and tokenization: (1) TagPrime [74], a sequence tagger priming words to input text to convey more task-specific information. (2) EEQA [42], a classification model utilizing label semantics by formulating event extraction as a question-answering task. (3) DyGIE++ [216], a multi-task end-to-end classification-based model using span graph propagation to aggregate local and global context. (4) OneIE [125], a joint-training model training a single model for various information extraction tasks. (5) XGear [85]. a language-agnostic generative EAE model which applies generative methods for EE. Since XGear is an EAE-only model, we combine it with the TagPrime ED model. To further improve the models, we train them with pseudo-generated data using a label projection model CLaP [162].

7.3.2 Baseline Models

As baselines for English-only ED, we consider zero-shot ED models (**Zero-shot**) that do not train on any supervised data and solely utilize the event definitions. We consider the following zero-shot models:

1. TE [133], a pre-trained model that uses event definitions to formulate ED as a textual entailment and question-answering task.

2. WSD [241] which encodes the contextualized trigger and event definitions jointly and uses a classification head atop for event detection.

3. ETypeClus [191], a pipeline that extracts salient predicate-object pairs and clusters the embeddings of these pairs in a spherical latent space.

We also consider transferring from existing datasets (**Transfer from Existing Datasets**) by training models on standard ED datasets like ACE [37] and MAVEN [220] without fine-tuning on epidemic ED data.

As stronger baselines, we also consider models utilizing epidemic ED data. Here, we consider models using few-shot target disease data without any model training (**No Training**) like:

1. Keyword [108], an epidemiological model utilizing curated event-specific keywords to detect events.

2. GPT-3.5 [18], a large-language model (LLM) using GPT-3.5-turbo with seven target disease in-context ED examples.

Finally, we consider super-strong baselines training ED models on limited 300 tweets for the target disease (**Trained on Target Epidemic**). Noting that these models are added for comparison, but they are practically infeasible for epidemic prediction, as it takes 4-6 weeks after the first infection to collect such target disease data.

Additionally, for wholistic EE evaluations, we also consider two other baselines of DivED [20], a Llama2-7B model fine-tuned on a diverse range of event definitions, and COVIDKB [257], an epidemiological work using a BERT classification model. Since the original output classes are different, we train it to simply classify tweets as epidemic-related or not.

7.3.3 Results

We present the results for our English-only ED models in Table 7.7. Firstly, none of the existing data transfer, zero-shot, or no training-based models perform well for our task, mainly owing to the domain shift of social media and unseen epidemic events. Overall, ED models trained on SPEED perform the best, thus demonstrating the capability of our ED framework to detect epidemic events for new diseases. Compared with models trained on the target epidemic, SPEED-trained models provide a gain of 10 F1 points for Monkeypox

Model	Monkeypox		Zika + Dengue	
	Tri-I	Tri-C	Tri-I	Tri-C
Zero-shot				
TE	16.70	12.11	12.69	9.06
WSD	22.04	4.35	27.93	5.85
ETypeClus	18.31	6.78	13.99	5.33
Transfer from Existing Datasets				
ACE - TagPrime	4.80	0	23.64	0
ACE - DEGREE	12.15	5.14	14.47	0
MAVEN - TagPrime	29.16	0	33.97	0
MAVEN - DEGREE	27.94	0	32.04	0
No Training				
Keyword	36.40	25.09	25.93	21.69
GPT-3.5	42.23	35.33	53.22	14.27
Trained on Target Epidemic				
BERT-QA	59.8	54.08	94.92	80.89
DEGREE	59.58	54.12	86.21	78.76
TagPrime	55.57	49.65	96.67	**84.43**
DyGIE++	55.83	50.31	73.24	65.65
Trained on SPEED (Our Framework)				
BERT-QA	**67.38**	**64.17**	**96.77**	81.97
DEGREE	62.95	61.45	88.52	77.69
TagPrime	64.71	61.92	95.24	75.54
DyGIE++	62.76	59.82	91.8	80.34

Table 7.7: Evaluating ED models trained on SPEED for detecting events for new epidemics of Monkeypox, Zika, and Dengue in terms of F1 scores.

and at par performance for Zika and Dengue. This outcome is particularly encouraging, as it demonstrates the resilience of our framework, making it highly applicable during the early stages of an epidemic, when minimal to no epidemic-specific data is accessible.

We present our per-disease per-language results in Table 7.8 and 7.9. The GPT-based LLM baseline performs better in English but exhibits poor performance across other languages. On the other hand, we observe stronger performance by the supervised baselines trained on our SPEED/SPEED++ dataset, with TagPrime providing the best overall average performance. We also note how CLaP further improves performance by 2-3 F1 points in the cross-lingual

Model	COVID			MPox	Zika + Dengue				Average
	hi	jp	es	en	en	hi	jp	es	
Baseline Models									
ACE - TagPrime	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
DivED*‡	0.0	0.0	27.0	36.7	47.7	4.4	1.3	12.8	16.2
Keyword⋆	15.2	28.6	26.3	41.3	39.3	18.6	**39.7**	20.6	28.7
COVIDKB†	45.2	42.3	24.2	18.5	45.5	47.5	34.0	34.6	36.5
GPT-3.5-turbo⋆	35.7	36.4	43.2	46.4	56.6	45.5	29.0	39.6	41.6
Trained on SPEED++ (Our Framework)									
TagPrime	60.1	35.3	62.0	**70.2**	**66.7**	65.0	27.2	**49.9**	<u>54.6</u>
TagPrime + XGear	60.1	35.3	62.0	**70.2**	**66.7**	65.0	27.2	**49.9**	54.6
BERT-QA	54.7	21.2	60.6	66.1	63.0	50.8	4.6	41.9	45.4
DyGIE++	61.0	38.2	61.7	67.4	64.1	61.4	27.5	45.6	53.4
OneIE	**61.9**	12.0	44.5	68.8	**66.7**	61.9	12.0	44.5	46.5
TagPrime + CLaP	58.6	**48.4**	**62.6**	**70.2**	**66.7**	**65.2**	39.2	49.7	**57.6**

Table 7.8: Benchmarking Event Classification (EC) of EE models trained on SPEED/SPEED++ for extracting event information in the cross-lingual cross-disease setting. Here, hi = Hindi, jp = Japanese, es = Spanish, and en = English. ⋆Numbers are higher due to string matching evaluation. †Binary classification evaluation.

setting, especially for character-based language Japanese.

7.4 Epidemic Prediction through Epidemic EE

As a practical validation of the utility of our framework, we evaluate whether SPEED-trained models are capable of providing early warnings for an unknown epidemic. More specifically, we aggregate the extracted event mentions by our framework over a time period and report any sharp increase in the rolling average of detected events as an epidemic warning. For evaluation, we compare it with the actual number of disease infections reported in the same time period. Naturally, the earlier we provide an epidemic warning, the better the framework is deemed.

‡English-based Llama performs poorly multilingually.

Model	COVID			MPox	Zika + Dengue				Average
	hi	jp	es	en	en	hi	jp	es	
Baseline Models									
ACE - TagPrime	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GPT-3.5-turbo*	14.5	15.0	16.8	24.0	33.0	20.0	11.0	15.1	18.7
Trained on SPEED++ (Our Framework)									
TagPrime	**39.0**	7.9	37.1	**45.1**	**48.5**	**40.7**	7.5	**28.2**	31.8
TagPrime + XGear	27.1	9.7	36.0	42.0	45.8	31.9	8.8	26.6	28.5
BERT-QA	33.9	4.0	28.1	39.0	45.5	31.1	0.8	24.1	25.8
DyGIE++	35.7	2.0	**39.1**	39.3	46.4	32.2	0.4	22.7	27.2
OneIE	34.3	11.4	37.3	42.6	47.9	38.5	5.0	25.4	30.3
TagPrime + CLaP	32.9	**19.1**	37.7	**45.1**	**48.5**	40.6	**18.8**	28.1	**33.9**

Table 7.9: Benchmarking Argument Classification (AC) of EE models trained on SPEED/SPEED++ for extracting event information in the cross-lingual cross-disease setting. Here, hi = Hindi, jp = Japanese, es = Spanish, and en = English. *Numbers are higher due to string matching evaluation. †Binary classification evaluation.

English-based Monkeypox Prediction For the first evaluation, we chose Monkeypox as the unseen disease and its outbreak from May 11 to Nov 11, 2022, as the unknown epidemic period. We utilize our English-only SPEED models to extract epidemic events and provide epidemic warnings. As our baseline models, we consider:

1. Keyword [108], an epidemiological model utilizing curated event-specific keywords to detect events.
2. Data transfer model trained on the standard ED dataset - MAVEN [220].
3. Epidemic-based classification model trained on the COVID-KB [257] dataset.
4. Unsupervised clustering model ETypeClus [191] utilizing a pipeline that extracts salient predicate-object pairs and clusters the embeddings of these pairs in a spherical latent space.

We report the number of epidemic events extracted by the BERT-QA trained on SPEED, along with the actual number of Monkeypox cases reported in the US [7] from May 11 to Nov

[7]As reported by CDC at https://www.cdc.gov/poxvirus/mpox/response/2022/mpx-trends.html

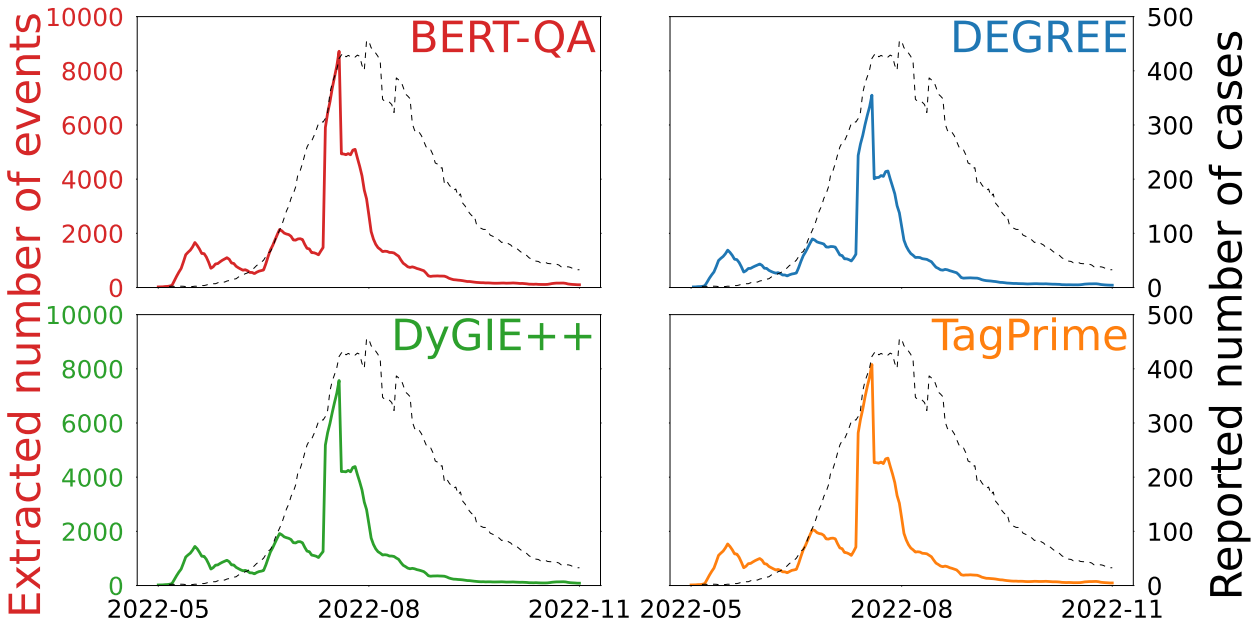

Figure 7.8: Number of reported Monkeypox cases and the number of extracted events from our four trained models from May 11 to Nov 11, 2022.

11, 2022, in Figure 7.2. As indicated by the arrows, our model could potentially provide two sets of early warnings around May 23 (9 weeks earlier, when first cases were detected) and June 29 (4 weeks earlier, when cases started rising) before the outbreak reached its peak around July 30. Comparatively, the MAVEN-trained model fails completely, while the other models' trends are too weak to provide any epidemic warnings. In fact, all our trained ED models are capable of providing these early signals as shown in Figure 7.8. This robust outcome underscores the practical utility of our framework to provide early epidemic warnings and ensure better preparedness for any potential epidemic.

Zero-shot Chinese-based COVID-19 Prediction We examine the earliest stages of the COVID-19 pandemic, analyzing Chinese social media posts from Dec 16, 2019, to Jan 21, 2020, using Weibo data from Hu et al. [81]. Using our trained TagPrime model, we infer on Chinese in a *zero-shot fashion* (i.e., without prior training on Chinese). We aggregate the 7-day rolling average of our extracted event mentions across time and report any sharp increases as epidemic warnings, as illustrated in Figure 7.1. Since case reporting had not

Chinese Posts	Translation
武汉华南海鲜市场出现 [infect]多个不明原因肺炎病例，请同道们提高警惕，早期发现，早期隔离 [prevent]	Multiple cases of pneumonia of unknown origin have appeared [infect] in Wuhan's Huanan Seafood Market. Please be more vigilant, detect and isolate [prevent] them as early as possible.
近日，武汉进入流感高发期 [spread]，多家医院感冒 [symptom]发烧的患者数量猛增。	Recently, Wuhan has entered a period of high influenza incidence [spread], and the number of patients with colds [symptom] and fevers in many hospitals has increased sharply.

Table 7.10: Sample Weibo posts in Chinese with their translations identified by SPEED/SPEED++ framework as epidemic-related from late December 2019. Event types and their trigger words are marked in blue.

begun, actual COVID-19 case numbers are unavailable for this period. Instead, we also plot the total number of Weibo posts and active users [64]. Additionally, we compare trends with baselines from COVIDKB [257] and a keyword-based approach [108].

As demonstrated in the figure, we see how our SPEED/SPEED++ framework provides epidemic warnings three weeks before the global tracking of infection cases began. While the keyword-based method also provides some signals, they are relatively weaker. Furthermore, Table 7.8 shows that the keyword baseline performs worse for morphologically richer languages, making it less robust. Additionally, the number of posts and active users do not provide any epidemic-related signals. For further validation, we present sample event mentions extracted by our framework in Table 7.10. In the bottom timeline[8] in Figure 7.1, we demonstrate the efficacy of our framework as it provided epidemic warnings 6 weeks before the "COVID-19" term was coined and used in social media. Overall, we show how our framework can provide early epidemic warnings multilingually without relying on any target language data, making it suitable for global deployment.

Global Epidemic Monitoring Validating our framework for each language is resource-intensive and infeasible. Instead, we perform a preliminary study of the broad language coverage of our framework by demonstrating its capability to detect COVID-related events

[8] https://www.cdc.gov/museum/timeline/covid19.html

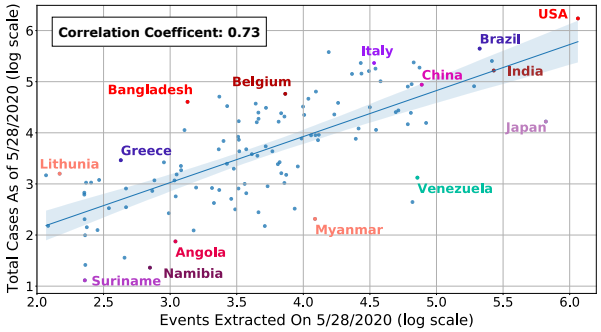

Figure 7.9: Number of extracted events plotted against the number of reported cases for each country. Both of them are in log scale.

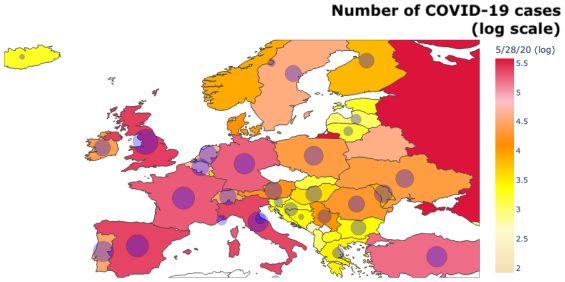

Figure 7.10: Geographical distribution of the number of reported COVID-19 cases as of May 28, 2020 in Europe. Red depicts more spread, and yellow/white indicates less spread. The blue dots indicate the events extracted by our model, and their size depicts the number of epidemic events for the specific country (log scale).

across 65 languages. We analyze tweets across all languages from a single day (May 28, 2020) and extract epidemic events using our framework. Utilizing user locations, we map the tweets from these languages to 117 countries. For reference, we plot the extracted events for each country with the actual number of reported COVID-19 cases[9] in Figure 7.9. Countries with significant COVID-19 spread appear in the top-right, while some outliers are also shown in the figure. Our framework achieves a healthy correlation of 0.73 with the actual reported cases, indicating strong performance across a broad range of languages.

We further extend these plots geographically in Figure 7.10. Each country is color-coded according to the number of COVID cases, with lighter shades indicating fewer cases while darker shades indicate a massive spread. The extracted number of events from the country-mapped tweets by our framework is plotted as translucent circles. Bigger dots indicate

[9] https://www.worldometers.info/coronavirus/

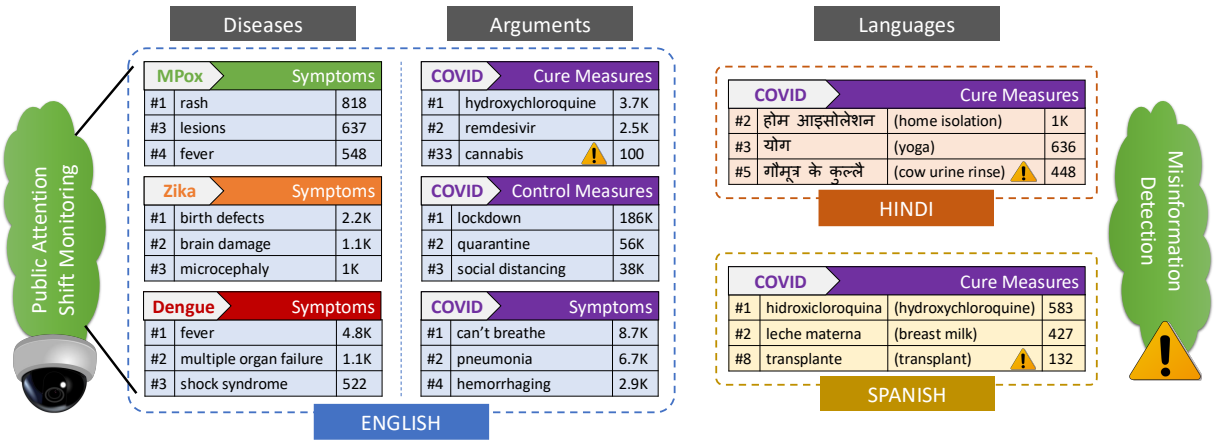

Figure 7.11: Information Assimilation Bulletin as extracted by our SPEED/SPEED++ framework and agglutinatively clustered. The first column represents different diseases, the second column represents different argument roles, and the third column represents the different languages. We also highlight the utility of these bulletins for two applications of Public Attention Shift Monitoring and Misinformation Detection.

more events extracted for the given country. In this plot, we observe extracted events and COVID-19 spread more in Western European countries like the United Kingdom, France, Spain, Italy, and Germany, while fewer events spread in Eastern European countries.

7.5 Epidemic Information Aggregation using Epidemic EE

Our framework possesses a strong EAE capability to extract detailed information about epidemic events such as *symptoms*, *preventive measures*, *cure measures*, etc. Aggregating such information from millions of social media posts can provide insights into public opinions regarding various epidemic aspects. To this end, we develop an information aggregation system for community epidemic discussions. Specifically, we use our framework to extract arguments for various event roles, project them into a representative space, and merge similar arguments using agglomerative clustering. The final arguments and their counts for different diseases and languages, extracted from 6M tweets, are presented as a bulletin in Figure 7.11.

Tweet	Translation
HINDI - COVID-19 - CURE MEASURES	
कोरोना के खिलाफ एक योद्धा बन के योग से कोरोना को हराना है।	Become a warrior against Corona and defeat it through yoga
संबित कोरोना पॉजिटिव, हालत नाजुक आत्मनिर्भर अभियान के तहत गौमूत्र के कुल्ले करवाकर उपचार किया जा रहा है।	Sambit is corona positive, condition is critical, and is being treated by gargling with cow urine under the self-reliant campaign.
संबित जी के कोरोना प्रभावित होने की खबर मिली। जानकर दुःख हुआ, लेकिन वे अस्पताल में स्वास्थ्य लाभ ले रहे हैं, और अब बहुत बेहतर है	Received the news of Sambit ji being affected by Corona! Sorry to hear, but he's recovering in the hospital, and is much better now.
SPANISH - COVID-19 - CURE MEASURES	
Científicos rusos sugieren que una proteína presente en la leche materna puede ser clave en la lucha contra el covid-19 (url)	Russian scientists suggest that a protein present in breast milk may be key in the fight against covid-19 (url)
Este transplante de pulmones a un paciente con COVID-19 es una operación realizada hasta ahora sólo en China y que por primera vez se lleva a cabo en Europa	This lung transplant to a patient with COVID-19 is an operation carried out so far only in China and is being carried out for the first time in Europe.
el mundo acumula más evidencia de la efectividad de la ivermectina para el tratamiento en casa de pacientes con estadios leves de #(COVID)-19	the world accumulates more evidence of the effectiveness of ivermectin for the home treatment of patients with mild stages of #(COVID)-19

Table 7.11: Illustration of some actual tweets in Hindi and Spanish mentioning various cure measures related to COVID-19. The terms related to *cure measures* extracted by our framework are highlighted in red.

Our framework effectively extracts various arguments for COVID-19 in English (middle column of Figure 7.11), including cure measures such as *hydroxychloroquine* and *remdesivir*, control measures like *lockdown* and *quarantine*, and symptoms such as *pneumonia*. Additionally, this capability extends to other diseases, such as Monkeypox, Zika, and Dengue (left column), and across languages, including Hindi and Spanish (right column). This condensed information is crucial for Public Attention Shift Monitoring, aiding policymakers in devising better control measures [128]. We demonstrate this in the form of extracted symptoms such as *rashes* and *lesions* for Monkeypox and *fever* and *shock syndrome* for Dengue (left column). Simultaneously, our framework can assist in **Misinformation Detection** [138]. Shown as

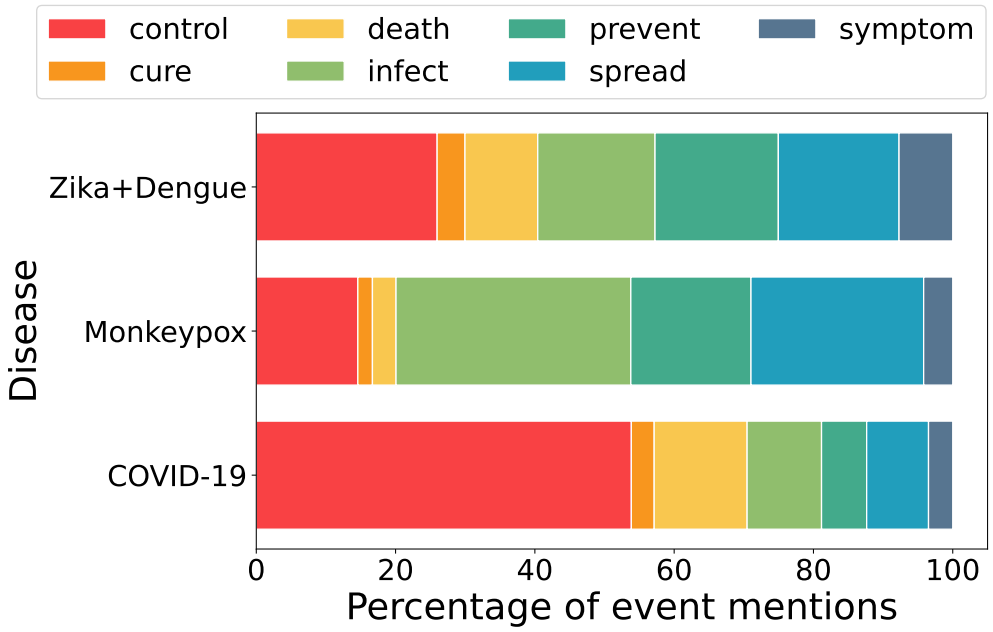

Figure 7.12: Disease profiles of public opinions generated by plotting the percentage of extracted event mentions by our SPEED/SPEED++ framework for COVID-19, Monkeypox, and Zika.

caution signs in the figure, we highlight various potential COVID-19 cure misinformation, such as *cow urine rinse*, *cannabis*, and *transplants* extracted across languages by our framework. As evidence, we also provide some of the actual tweets in Hindi and Spanish flagged by our framework for mentioning these terms in Table 7.11.

Event-based Disease Profiling Our framework offers the additional utility of generating event-based disease profiles using public sentiments. These disease profiles can be generated by plotting the percentage of mentions per event type extracted by our framework. Using 500k tweets, we depict the profiles for COVID, Monkeypox, and Zika+Dengue in Figure 7.12.

Distinctive profiles emerge for each disease; COVID-19 majorly comprises *control*, Monkeypox exhibits a bias toward *infect* and *spread*, while Zika+Dengue emphasizes *control* and *death*. These trends align with the higher fatality rate of Zika and Dengue [156], recent discoveries of transmission routes of Monkeypox [100], and the need for mass public control measures for the COVID pandemic [62]. Relatively, Monkeypox also shows low mentions for

	Disease	Infect Event Example	Symptom Event Example
Keyword-based	COVID-19	Three students infected with **COVID-19**	COVID-19 symptoms include **fever**, **cough**, ...
Keyword-based	Monkeypox	How do you catch **Monkeypox**?	Monkeypox may cause **rashes** and **itching** ...
SPEED (Ours)	COVID-19	Three students **infected** with COVID-19	COVID-19 **symptoms** include fever, cough, ...
SPEED (Ours)	Monkeypox	How do you **catch** Monkeypox?	Monkeypox may **cause** rashes and itching ...

Table 7.12: Qualitative analysis for annotation difference between previous keyword-based epidemiological datasets [27, 108] and SPEED's Event Detection based annotation schema. Our annotation schema is less disease-specific and thus, better generalizable to a wide range of diseases.

death, *cure* - which aligns with low fatality and no available cure for Monkeypox [97]. Overall, these profiles can provide policymakers with valuable insights about new and unknown outbreaks to implement more informed and effective interventions.

Why does SPEED generalize? We provide a qualitative analysis of why COVID-based SPEED helps detect epidemic events for other unseen diseases compared to previous epidemiological works [27, 108] and attribute it to the difference in the task formulation and annotation schema. We demonstrate this difference (highlighted in **bold**) through illustrative examples for *Infect* and *Symptom* events in Table 7.12. As is evident, keyword-based modeling requires annotating highly precise but disease-specific keywords like *COVID-19*, *fever*, etc. On the other hand, our ED annotation formulation emphasizes the annotation of disease-agnostic triggers like *infected*, *symptoms*, etc. This provides SPEED and our framework superior generalizability without new annotations to unseen diseases.

7.6 Conclusion

Overall, in our work for this chapter, we pioneer the creation of the first multilingual Event Extraction (EE) framework for application in epidemic prediction and preparedness. To

this end, we create a multilingual EE benchmarking dataset SPEED/SPEED++ comprising 5K tweets spanning four languages and four diseases. To realistically deploy our models, we develop zero-shot cross-lingual cross-disease models and demonstrate their capability to extract events for 65 languages spanning 117 countries. We prove the effectiveness of our model by providing early epidemic warnings for Monkeypox from English Twitter and COVID-19 from Chinese Weibo posts in a zero-shot manner. We also show evidence that our framework can be utilized as an information aggregation system aiding in misinformation detection and public attention monitoring. In conclusion, we provide a strong application of multilingual EE for global epidemic preparedness.

Chapter 8

Future Work

In my dissertation, I have proposed various methodologies utilizing synthetic data curation and multi-agent frameworks that significantly improve the generalization and application of Event Extraction. In this chapter, we outline broad directions for future work to further expand the horizons of Universal Event Extraction, categorized across its two foundational pillars: advancing open-domain generalizability and broadening real-world applicability.

8.1 Towards Adaptive Event Extraction

Current event extraction frameworks, including the core methodologies developed in this dissertation, largely operate under the assumption of a static, predefined event ontology. Deploying these systems "in the wild", however, exposes them to a constantly shifting information landscape where novel event types – such as unprecedented cyber-threats or emerging epidemiological phenomena – frequently arise. A critical future direction for advancing generalizability is the development of autonomous schema evolution mechanisms that move beyond fixed taxonomies. Drawing inspiration from open information extraction paradigms [244] and recent advances in generative schema induction [118], future multi-agent systems could dynamically detect out-of-ontology triggers and collaboratively propose new event nodes.

To support this dynamic ontology without suffering from catastrophic forgetting, the extraction framework must be coupled with self-improving data loops. Leveraging inverse data generation pipelines akin to the SNaRe framework introduced in this thesis, an adaptive system could actively monitor its own extraction confidence on live text streams. Upon identifying low-confidence zones or newly proposed event classes, the framework would trigger a targeted synthesis process to generate synthetic "booster" datasets for those specific edge cases. This closed-loop paradigm – where the model identifies its own semantic gaps and orchestrates synthetic data generation for continuous fine-tuning – mirrors recent successes in self-instruction and iterative alignment [222, 83]. Ultimately, this would transform event extraction from rigid structural snapshots into fluid representations that organically expand their semantic boundaries in response to real-time, streaming data.

8.2 Improving Discriminative Reasoning in LLMs

While large language models demonstrate profound generative fluency, their discriminative generalization remains a significant bottleneck in structural tasks like event extraction. As the cardinality of event ontologies increases, models often struggle with boundary resolution, conflating semantically proximate event types. Recent advancements have demonstrated that scaling inference-time compute – also called "System-2" reasoning [240, 195] – can substantially elevate performance on complex reasoning tasks. However, applying unconstrained test-time scaling to event extraction is inefficient and often insufficient without targeted structural constraints.

A critical future direction to mitigate the performance degradation associated with high-density schemas is tackling scale through advanced, post-trained multi-agent collaboration. Rather than relying on a single LLM to parse an entire massive ontology, the extraction workload can be distributed across decentralized agent swarms [230, 256]. In this paradigm, distinct "expert" sub-swarms are dynamically instantiated to specialize in specific semantic

subspaces – such as corporate, legal, or geopolitical events. Although orchestrating a swarm of specialized LLM agents introduces higher computational inference costs, it intrinsically circumvents the context-window dilution and "lost in the middle" degradation that direly affects single-pass prompting. By synthesizing the localized, discriminative intelligence of these expert swarms through a centralized orchestration agent, we can construct highly robust event graphs that maintain high fidelity even as the target application scales.

8.3 Event-Centric World Modeling

Transitioning to the applications of Universal EE, as large language models increasingly generate long-form text and complex narratives, assessing their true semantic comprehension remains a significant challenge. A vital future direction is leveraging extracted event graphs as a structural gold standard for event-centric evaluation. By formalizing text into discrete event skeletons, we can systematically benchmark LLM-generated outputs directly against human narratives. This approach shifts the evaluation paradigm from surface-level fluency to structural fidelity, allowing us to rigorously quantify narrative depth, event density, and causal consistency. Such methodologies will expose whether models merely mimic statistical distributions of language or genuinely capture the underlying mechanics and logic of a complex story [94, 157].

Beyond passive evaluation, these unified event graphs offer a foundational substrate for active world modeling and AI simulation. By mapping the intricate causal edges between real-world occurrences, these graphs can function as interactive, counterfactual sandboxes. Future research could systematically perturb specific seed events (e.g., simulating a sudden port closure or an unpredicted epidemiological shift) and computationally propagate these changes through the graph to generate a counterfactual world state. Integrating these structured environments enables the rigorous training and auditing of AI agents in hypothetical, high-stakes scenarios. Overall, this positions universal event extraction not just as a retrospective

analytical tool, but as a critical foresight engine, equipping policy-making agents with the causal grounding necessary to safely navigate simulated futures [167, 239].

8.4 Scaling EE for High-Stakes Clinical and Financial Applications

While existing literature has explored the utility of Event Extraction in clinical domains for drug discovery [219], adverse case detection [178], and workflow optimizations [78], as well as in the financial sector [46, 63, 258], a significant opportunity remains to deploy modern, generative EE models for more expansive, end-to-end applications in these fields.

This cross-disciplinary direction is purely application-driven, aiming to identify practical bottlenecks in specialized disciplines and provide robust solutions utilizing Universal EE as the core computational engine. Potential high-impact applications include:

- **Financial News Analysis:** Tracking, summarizing, and linking global events to specific firms and corporations to power automated trading, risk management, and regulatory compliance.

- **Earnings Call Diagnostics:** Extracting structured event timelines from corporate communications to help investors and analysts systematically evaluate company performance and financial health.

- **Clinical Documentation:** Deploying on-the-fly event extraction to summarize the vast amounts of unstructured clinical data generated daily, thereby improving administrative efficiency and patient record compliance.

- **Clinical Decision Support:** Providing succinct, historically grounded event trajectories from patient profiles to assist healthcare providers in making informed diagnoses and treatment decisions.

Overall, pursuing these localized implementations will critically bolster the applicability of event extraction across diverse, high-stakes industries.

Bibliography

[1] David Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba Alabi, Shamsuddeen Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, Amelia Taylor, Fatoumata Kabore, Chris Chinenye Emezue, Anuoluwapo Aremu, Perez Ogayo, Catherine Gitau, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Mboning Tchiaze Elvis, Tajuddeen Gwadabe, Tosin Adewumi, Orevaoghene Ahia, Joyce Nakatumba-Nabende, Neo Lerato Mokono, Ignatius Ezeani, Chiamaka Chukwuneke, Mofetoluwa Oluwaseun Adeyemi, Gilles Quentin Hacheme, Idris Abdulmumin, Odunayo Ogundepo, Oreen Yousuf, Tatiana Moteu, and Dietrich Klakow. MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4488–4508, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.298. URL https://aclanthology.org/2022.emnlp-main.298.

[2] Shantanu Agarwal, Steven Fincke, Chris Jenkins, Scott Miller, and Elizabeth Boschee. Impact of subword pooling strategy on cross-lingual event detection. *CoRR*, abs/2302.11365, 2023. doi: 10.48550/arXiv.2302.11365. URL https://doi.org/10.48550/arXiv.2302.11365.

[3] Jacqueline Aguilar, Charley Beller, Paul McNamee, Benjamin Van Durme, Stephanie Strassel, Zhiyi Song, and Joe Ellis. A comparison of the events and relations across ACE, ERE, TAC-KBP, and FrameNet annotation standards. In Teruko Mitamura, Eduard Hovy, and Martha Palmer, editors, *Proceedings of the Second Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 45–53, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-2907. URL https://aclanthology.org/W14-2907.

[4] Wasi Ahmad, Haoran Li, Kai-Wei Chang, and Yashar Mehdad. Syntax-augmented multilingual BERT for cross-lingual transfer. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4538–4554, Online, August 2021.

Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.350. URL https://aclanthology.org/2021.acl-long.350.

[5] Wasi Uddin Ahmad, Zhisong Zhang, Xuezhe Ma, Kai-Wei Chang, and Nanyun Peng. Cross-lingual dependency parsing with unlabeled auxiliary languages. In Mohit Bansal and Aline Villavicencio, editors, *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 372–382, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/K19-1035. URL https://aclanthology.org/K19-1035.

[6] Wasi Uddin Ahmad, Nanyun Peng, and Kai-Wei Chang. GATE: graph attention transformer encoder for cross-lingual relation and event extraction. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 12462–12470. AAAI Press, 2021. doi: 10.1609/aaai.v35i14.17478. URL https://doi.org/10.1609/aaai.v35i14.17478.

[7] David Ahn. The stages of event extraction. In Branimir Boguraev, Rafael Muñoz, and James Pustejovsky, editors, *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, pages 1–8, Sydney, Australia, July 2006. Association for Computational Linguistics. URL https://aclanthology.org/W06-0901.

[8] Alan Akbik, Laura Chiticariu, Marina Danilevsky, Yunyao Li, Shivakumar Vaithyanathan, and Huaiyu Zhu. Generating high quality proposition Banks for multilingual semantic role labeling. In Chengqing Zong and Michael Strube, editors, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 397–407, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-1039. URL https://aclanthology.org/P15-1039.

[9] Arthur Amalvy, Vincent Labatut, and Richard Dufour. Learning to rank context for named entity recognition using a synthetic dataset. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10372–10382, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.642. URL https://aclanthology.org/2023.emnlp-main.642/.

[10] Maryam Aminian, Mohammad Sadegh Rasooli, and Mona Diab. Transferring semantic roles using translation and syntactic information. In Greg Kondrak and Taro Watanabe, editors, *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 13–19, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing. URL https://aclanthology.org/I17-2003.

[11] Maryam Aminian, Mohammad Sadegh Rasooli, and Mona Diab. Cross-lingual transfer of semantic roles: From raw text to semantic roles. In Simon Dobnik, Stergios Chatzikyriakidis, and Vera Demberg, editors, *Proceedings of the 13th International Conference*

on Computational Semantics - Long Papers, pages 200–210, Gothenburg, Sweden, May 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-0417. URL https://aclanthology.org/W19-0417.

[12] Abhijeet Awasthi, Nitish Gupta, Bidisha Samanta, Shachi Dave, Sunita Sarawagi, and Partha Talukdar. Bootstrapping multilingual semantic parsers using large language models. In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2455–2467, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.180. URL https://aclanthology.org/2023.eacl-main.180.

[13] Shane Babcock, John Beverley, Lindsay G. Cowell, and Barry Smith. The infectious disease ontology in the age of COVID-19. *J. Biomed. Semant.*, 12(1):13, 2021. doi: 10.1186/s13326-021-00245-1. URL https://doi.org/10.1186/s13326-021-00245-1.

[14] Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 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, August 1998. Association for Computational Linguistics. doi: 10.3115/980845.980860. URL https://aclanthology.org/P98-1013.

[15] Debangshu Banerjee, Tarun Suresh, Shubham Ugare, Sasa Misailovic, and Gagandeep Singh. CRANE: reasoning with constrained LLM generation. *CoRR*, abs/2502.09061, 2025. doi: 10.48550/ARXIV.2502.09061. URL https://doi.org/10.48550/arXiv.2502.09061.

[16] Jonathan Berant, Vivek Srikumar, Pei-Chun Chen, Abby Vander Linden, Brittany Harding, Brad Huang, Peter Clark, and Christopher D. Manning. Modeling biological processes for reading comprehension. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1499–1510, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1159. URL https://aclanthology.org/D14-1159.

[17] Mihaela A. Bornea, Lin Pan, Sara Rosenthal, Radu Florian, and Avirup Sil. Multilingual transfer learning for QA using translation as data augmentation. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 12583–12591. AAAI Press, 2021. doi: 10.1609/aaai.v35i14.17491. URL https://doi.org/10.1609/aaai.v35i14.17491.

[18] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon

Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

[19] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020. URL https://arxiv.org/abs/2005.14165.

[20] Zefan Cai, Po-Nien Kung, Ashima Suvarna, Mingyu Ma, Hritik Bansal, Baobao Chang, P. Jeffrey Brantingham, Wei Wang, and Nanyun Peng. Improving event definition following for zero-shot event detection. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2842–2863, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.157. URL https://aclanthology.org/2024.acl-long.157/.

[21] Pengfei Cao, Zhuoran Jin, Yubo Chen, Kang Liu, and Jun Zhao. Zero-shot cross-lingual event argument extraction with language-oriented prefix-tuning. In Brian Williams, Yiling Chen, and Jennifer Neville, editors, *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 12589–12597. AAAI Press, 2023. doi: 10.1609/aaai.v37i11.26482. URL https://doi.org/10.1609/aaai.v37i11.26482.

[22] Feng Chen and Yujian Feng. Chain-of-thought prompt distillation for multimodal named entity and multimodal relation extraction. *arXiv preprint arXiv:2306.14122*, 2023.

[23] Yang Chen and Alan Ritter. Model selection for cross-lingual transfer. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5675–5687, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.459. URL https://aclanthology.org/2021.emnlp-main.459.

[24] Yang Chen, Chao Jiang, Alan Ritter, and Wei Xu. Frustratingly easy label projection for cross-lingual transfer. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5775–5796, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.357. URL https://aclanthology.org/2023.findings-acl.357.

[25] Yew Ken Chia, Lidong Bing, Soujanya Poria, and Luo Si. RelationPrompt: Leveraging prompts to generate synthetic data for zero-shot relation triplet extraction. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 45–57, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.5. URL https://aclanthology.org/2022.findings-acl.5/.

[26] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24:240:1–240:113, 2023. URL http://jmlr.org/papers/v24/22-1144.html.

[27] Nigel Collier, Son Doan, Ai Kawazoe, Reiko Matsuda Goodwin, Mike Conway, Yoshio Tateno, Hung Quoc Ngo, Dinh Dien, Asanee Kawtrakul, Koichi Takeuchi, Mika Shigematsu, and Kiyosu Taniguchi. Biocaster: detecting public health rumors with a web-based text mining system. *Bioinform.*, 24(24):2940–2941, 2008. doi: 10.1093/bioinformatics/btn534. URL https://doi.org/10.1093/bioinformatics/btn534.

[28] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116, 2019. URL http://arxiv.org/abs/1911.02116.

[29] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July

2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747.

[30] Nicola De Cao, Ledell Wu, Kashyap Popat, Mikel Artetxe, Naman Goyal, Mikhail Plekhanov, Luke Zettlemoyer, Nicola Cancedda, Sebastian Riedel, and Fabio Petroni. Multilingual autoregressive entity linking. *Transactions of the Association for Computational Linguistics*, 10:274–290, 2022. doi: 10.1162/tacl_a_00460. URL https://aclanthology.org/2022.tacl-1.16.

[31] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, and S. S. Li. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948, 2025. doi: 10.48550/ARXIV.2501.12948. URL https://doi.org/10.48550/arXiv.2501.12948.

[32] Shumin Deng, Ningyu Zhang, Jiaojian Kang, Yichi Zhang, Wei Zhang, and Huajun Chen. Meta-learning with dynamic-memory-based prototypical network for few-shot event detection. *CoRR*, abs/1910.11621, 2019. URL http://arxiv.org/abs/1910.11621.

[33] Shumin Deng, Ningyu Zhang, Jiaojian Kang, Yichi Zhang, Wei Zhang, and Huajun Chen. Meta-learning with dynamic-memory-based prototypical network for few-shot event detection. In *The Thirteenth ACM International Conference on Web Search and Data Mining (WSDM)*, 2020. URL https://doi.org/10.1145/3336191.3371796.

[34] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

[35] Guilherme Dias. Tweets dataset on Zika virus. 5 2020. doi: 10.6084/m9.figshare.

12301400.v1. URL https://figshare.com/articles/dataset/Tweets_dataset_on_Zika_virus/12301400.

[36] Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. Is GPT-3 a good data annotator? In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11173–11195, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.626. URL https://aclanthology.org/2023.acl-long.626/.

[37] George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. The automatic content extraction (ACE) program – tasks, data, and evaluation. In Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa, and Raquel Silva, editors, *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May 2004. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2004/pdf/5.pdf.

[38] Rezarta Islamaj Dogan, Robert Leaman, and Zhiyong Lu. NCBI disease corpus: A resource for disease name recognition and concept normalization. *J. Biomed. Informatics*, 47:1–10, 2014. doi: 10.1016/j.jbi.2013.12.006. URL https://doi.org/10.1016/j.jbi.2013.12.006.

[39] Zi-Yi Dou and Graham Neubig. Word alignment by fine-tuning embeddings on parallel corpora. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.181. URL https://aclanthology.org/2021.eacl-main.181.

[40] Timothy J. Doyle, Haobo Ma, Samuel L. Groseclose, and Richard S. Hopkins. Phskb: A knowledgebase to support notifiable disease surveillance. *BMC Medical Informatics Decis. Mak.*, 5:27, 2005. doi: 10.1186/1472-6947-5-27. URL https://doi.org/10.1186/1472-6947-5-27.

[41] Mian Du, Peter von Etter, Mikhail Kopotev, Mikhail Novikov, Natalia Tarbeeva, and Roman Yangarber. Building support tools for russian-language information extraction. In Ivan Habernal and Václav Matousek, editors, *Text, Speech and Dialogue - 14th International Conference, TSD 2011, Pilsen, Czech Republic, September 1-5, 2011. Proceedings*, volume 6836 of *Lecture Notes in Computer Science*, pages 380–387. Springer, 2011. doi: 10.1007/978-3-642-23538-2_48. URL https://doi.org/10.1007/978-3-642-23538-2_48.

[42] Xinya Du and Claire Cardie. Event extraction by answering (almost) natural questions. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 671–683, Online, November 2020. Association for Computational Linguistics. doi: 10.

18653/v1/2020.emnlp-main.49. URL https://aclanthology.org/2020.emnlp-main.49.

[43] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. The llama 3 herd of models. *CoRR*, abs/2407.21783, 2024. doi: 10.48550/ARXIV.2407.21783. URL https://doi.org/10.48550/arXiv.2407.21783.

[44] Chris Dyer, Victor Chahuneau, and Noah A. Smith. A simple, fast, and effective reparameterization of IBM model 2. In Lucy Vanderwende, Hal Daumé III, and Katrin Kirchhoff, editors, *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL https://aclanthology.org/N13-1073.

[45] Seth Ebner, Patrick Xia, Ryan Culkin, Kyle Rawlins, and Benjamin Van Durme. Multi-sentence argument linking. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8057–8077, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.718. URL https://aclanthology.org/2020.acl-main.718.

[46] Liat Ein-Dor, Ariel Gera, Orith Toledo-Ronen, Alon Halfon, Benjamin Sznajder, Lena Dankin, Yonatan Bilu, Yoav Katz, and Noam Slonim. Financial event extraction using Wikipedia-based weak supervision. In Udo Hahn, Véronique Hoste, and Zhu Zhang, editors, *Proceedings of the Second Workshop on Economics and Natural Language Processing*, pages 10–15, Hong Kong, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5102. URL https://aclanthology.org/D19-5102.

[47] Joe Ellis, Jeremy Getman, and Stephanie M Strassel. 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, 2014.

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

[49] Gunther Eysenbach. Infodemiology: Tracking flu-related searches on the web for syndromic surveillance. In *AMIA 2006, American Medical Informatics Association Annual Symposium, Washington, DC, USA, November 11-15, 2006*. AMIA, 2006. URL https://knowledge.amia.org/amia-55142-a2006a-1.620145/t-001-1.623243/f-001-1.623244/a-049-1.623601/a-050-1.623598.

[50] Hao Fei, Meishan Zhang, and Donghong Ji. Cross-lingual semantic role labeling with high-quality translated training corpus. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7014–7026, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.627. URL https://aclanthology.org/2020.acl-main.627.

[51] Hao Fei, Shengqiong Wu, Jingye Li, Bobo Li, Fei Li, Libo Qin, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Lasuie: Unifying information extraction with latent adaptive structure-aware generative language model. *CoRR*, abs/2304.06248, 2023. doi: 10.48550/ARXIV.2304.06248. URL https://doi.org/10.48550/arXiv.2304.06248.

[52] Charles J Fillmore et al. 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, 1976.

[53] Steven Fincke, Shantanu Agarwal, Scott Miller, and Elizabeth Boschee. Language model priming for cross-lingual event extraction. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 10627–10635. AAAI Press, 2022. doi: 10.1609/aaai.v36i10.21307. URL https://doi.org/10.1609/aaai.v36i10.21307.

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

[55] Clark C. Freifeld, Kenneth D. Mandl, Ben Y. Reis, and John S. Brownstein. Model formulation: Healthmap: Global infectious disease monitoring through automated classification and visualization of internet media reports. *J. Am. Medical Informatics Assoc.*, 15(2):150–157, 2008. doi: 10.1197/jamia.M2544. URL https://doi.org/10.1197/jamia.M2544.

[56] Jun Gao, Huan Zhao, Changlong Yu, and Ruifeng Xu. Exploring the feasibility of chatgpt for event extraction. *CoRR*, abs/2303.03836, 2023. doi: 10.48550/ARXIV.2303.03836. URL https://doi.org/10.48550/arXiv.2303.03836.

[57] Jun Gao, Huan Zhao, Changlong Yu, and Ruifeng Xu. Exploring the feasibility of chatgpt for event extraction. *CoRR*, abs/2303.03836, 2023. doi: 10.48550/ARXIV.2303.03836. URL https://doi.org/10.48550/arXiv.2303.03836.

[58] Saibo Geng, Martin Josifoski, Maxime Peyrard, and Robert West. Grammar-constrained decoding for structured NLP tasks without finetuning. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10932–10952, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.674. URL https://aclanthology.org/2023.emnlp-main.674/.

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

[60] Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, 2009.

[61] Ralph Grishman and Beth Sundheim. Message Understanding Conference- 6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*, 1996. URL https://aclanthology.org/C96-1079.

[62] Hatice Rahmet Güner, İmran Hasanoğlu, and Firdevs Aktaş. Covid-19: Prevention and control measures in community. *Turkish Journal of medical sciences*, 50(9):571–577, 2020.

[63] Kaihao Guo, Tianpei Jiang, and Haipeng Zhang. Knowledge graph enhanced event extraction in financial documents. In Xintao Wu, Chris Jermaine, Li Xiong, Xiaohua Hu, Olivera Kotevska, Siyuan Lu, Weija Xu, Srinivas Aluru, Chengxiang Zhai, Eyhab Al-Masri, Zhiyuan Chen, and Jeff Saltz, editors, *2020 IEEE International Conference on Big Data (IEEE BigData 2020), Atlanta, GA, USA, December 10-13, 2020*, pages 1322–1329. IEEE, 2020. doi: 10.1109/BIGDATA50022.2020.9378471. URL https://doi.org/10.1109/BigData50022.2020.9378471.

[64] Shuhui Guo, Fan Fang, Tao Zhou, Wei Zhang, Qiang Guo, Rui bi Zeng, Xiaohong Chen, Jianguo Liu, and Xin Lu. Improving google flu trends for covid-19 estimates using weibo posts. *Data Science and Management*, 3:13 – 21, 2021. URL https://api.semanticscholar.org/CorpusID:235915214.

[65] Felix Hamborg, Corinna Breitinger, and Bela Gipp. Giveme5w1h: A universal system for extracting main events from news articles. In Özlem Özgöbek, Benjamin Kille, Jon Atle Gulla, and Andreas Lommatzsch, editors, *Proceedings of the 7th International Workshop on News Recommendation and Analytics in conjunction with 13th ACM Conference*

on Recommender Systems, INRA@RecSys 2019, Copenhagen, Denmark, September 20, 2019, volume 2554 of *CEUR Workshop Proceedings*, pages 35–43. CEUR-WS.org, 2019. URL https://ceur-ws.org/Vol-2554/paper_06.pdf.

[66] Rujun Han, I-Hung Hsu, Mu Yang, Aram Galstyan, Ralph Weischedel, and Nanyun Peng. Deep structured neural network for event temporal relation extraction. In Mohit Bansal and Aline Villavicencio, editors, *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 666–106, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/K19-1062. URL https://aclanthology.org/K19-1062.

[67] Rujun Han, I-Hung Hsu, Jiao Sun, Julia Baylon, Qiang Ning, Dan Roth, and Nanyun Peng. ESTER: A machine reading comprehension dataset for reasoning about event semantic relations. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7543–7559, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.597. URL https://aclanthology.org/2021.emnlp-main.597.

[68] Xuanli He, Islam Nassar, Jamie Kiros, Gholamreza Haffari, and Mohammad Norouzi. Generate, annotate, and learn: Generative models advance self-training and knowledge distillation. *CoRR*, abs/2106.06168, 2021. URL https://arxiv.org/abs/2106.06168.

[69] Yongqun He, Hong Yu, Edison Ong, Yang Wang, Yingtong Liu, Anthony Huffman, Hsin-Hui Huang, John Beverley, Asiyah Yu Lin, William D. Duncan, Sivaram Arabandi, Jiangan Xie, Junguk Hur, Xiaolin Yang, Luonan Chen, Gilbert S. Omenn, Brian D. Athey, and Barry Smith. CIDO: the community-based coronavirus infectious disease ontology. In Janna Hastings and Frank Loebe, editors, *Proceedings of the 11th International Conference on Biomedical Ontologies (ICBO) joint with the 10th Workshop on Ontologies and Data in Life Sciences (ODLS) and part of the Bolzano Summer of Knowledge (BoSK 2020), Virtual conference hosted in Bolzano, Italy, September 17, 2020*, volume 2807 of *CEUR Workshop Proceedings*, pages 1–10. CEUR-WS.org, 2020. URL https://ceur-ws.org/Vol-2807/paperE.pdf.

[70] Leonhard Hennig, Philippe Thomas, and Sebastian Möller. MultiTACRED: A multilingual version of the TAC relation extraction dataset. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3785–3801, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.210. URL https://aclanthology.org/2023.acl-long.210.

[71] David L Heymann, Guénaël R Rodier, et al. Hot spots in a wired world: Who surveillance of emerging and re-emerging infectious diseases. *The Lancet infectious diseases*, 1(5):345–353, 2001.

[72] Frederik Hogenboom, Flavius Frasincar, Uzay Kaymak, Franciska de Jong, and Emiel Caron. A survey of event extraction methods from text for decision support systems.

Decis. Support Syst., 85:12–22, 2016. doi: 10.1016/J.DSS.2016.02.006. URL https://doi.org/10.1016/j.dss.2016.02.006.

[73] I-Hung Hsu, Kuan-Hao Huang, Elizabeth Boschee, Scott Miller, Prem Natarajan, Kai-Wei Chang, and Nanyun Peng. DEGREE: A data-efficient generation-based event extraction model. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *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, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.138. URL https://aclanthology.org/2022.naacl-main.138.

[74] I-Hung Hsu, Kuan-Hao Huang, Shuning Zhang, Wenxin Cheng, Prem Natarajan, Kai-Wei Chang, and Nanyun Peng. TAGPRIME: A unified framework for relational structure extraction. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12917–12932, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.723. URL https://aclanthology.org/2023.acl-long.723.

[75] I-Hung Hsu, Avik Ray, Shubham Garg, Nanyun Peng, and Jing Huang. Code-switched text synthesis in unseen language pairs. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5137–5151, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.318. URL https://aclanthology.org/2023.findings-acl.318.

[76] I-Hung Hsu, Zhiyu Xie, Kuan-Hao Huang, Prem Natarajan, and Nanyun Peng. AMPERE: AMR-aware prefix for generation-based event argument extraction model. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10976–10993, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.615. URL https://aclanthology.org/2023.acl-long.615.

[77] I-Hung Hsu, Zifeng Wang, Long Le, Lesly Miculicich, Nanyun Peng, Chen-Yu Lee, and Tomas Pfister. CaLM: Contrasting large and small language models to verify grounded generation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12782–12803, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.759. URL https://aclanthology.org/2024.findings-acl.759/.

[78] William Hsu, Simon X. Han, Corey W. Arnold, Alex A. T. Bui, and Dieter R. Enzmann. A data-driven approach for quality assessment of radiologic interpretations. *J. Am. Medical Informatics Assoc.*, 23(e1):e152–e156, 2016. doi: 10.1093/JAMIA/OCV161. URL https://doi.org/10.1093/jamia/ocv161.

[79] Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *CoRR*, abs/2003.11080, 2020. URL https://arxiv.org/abs/2003.11080.

[80] Xuming Hu, Yong Jiang, Aiwei Liu, Zhongqiang Huang, Pengjun Xie, Fei Huang, Lijie Wen, and Philip S. Yu. Entity-to-text based data augmentation for various named entity recognition tasks. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9072–9087, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.578. URL https://aclanthology.org/2023.findings-acl.578/.

[81] Yong Hu, Heyan Huang, Anfan Chen, and Xian-Ling Mao. Weibo-COV: A large-scale COVID-19 social media dataset from Weibo. In Karin Verspoor, Kevin Bretonnel Cohen, Michael Conway, Berry de Bruijn, Mark Dredze, Rada Mihalcea, and Byron Wallace, editors, *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online, December 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.nlpcovid19-2.34. URL https://aclanthology.org/2020.nlpcovid19-2.34/.

[82] James Y. Huang, Bangzheng Li, Jiashu Xu, and Muhao Chen. Unified semantic typing with meaningful label inference. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2642–2654, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.190. URL https://aclanthology.org/2022.naacl-main.190.

[83] Jiaxin Huang, Shixiang Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. Large language models can self-improve. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1051–1068, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.67. URL https://aclanthology.org/2023.emnlp-main.67/.

[84] Kuan-Hao Huang, Wasi Ahmad, Nanyun Peng, and Kai-Wei Chang. Improving zero-shot cross-lingual transfer learning via robust training. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1684–1697, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.126. URL https://aclanthology.org/2021.emnlp-main.126.

[85] Kuan-Hao Huang, I-Hung Hsu, Prem Natarajan, Kai-Wei Chang, and Nanyun Peng. Multilingual generative language models for zero-shot cross-lingual event argument extraction. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors,

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 4633–4646, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.317. URL https://aclanthology.org/2022.acl-long.317.

[86] Kuan-Hao Huang, I-Hung Hsu, Tanmay Parekh, Zhiyu Xie, Zixuan Zhang, Prem Natarajan, Kai-Wei Chang, Nanyun Peng, and Heng Ji. TextEE: Benchmark, reevaluation, reflections, and future challenges in event extraction. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12804–12825, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.760. URL https://aclanthology.org/2024.findings-acl.760/.

[87] Lifu Huang, Heng Ji, Kyunghyun Cho, Ido Dagan, Sebastian Riedel, and Clare Voss. Zero-shot transfer learning for event extraction. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2160–2170, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1201. URL https://aclanthology.org/P18-1201.

[88] Pere-Lluís Huguet Cabot, Simone Tedeschi, Axel-Cyrille Ngonga Ngomo, and Roberto Navigli. RED$^{\text{fm}}$: a filtered and multilingual relation extraction dataset. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4326–4343, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.237. URL https://aclanthology.org/2023.acl-long.237/.

[89] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan,

Ian O'Connell, Ian Osband, Ignasi Clavera Gilaberte, and Ilge Akkaya. Openai o1 system card. *CoRR*, abs/2412.16720, 2024. doi: 10.48550/ARXIV.2412.16720. URL https://doi.org/10.48550/arXiv.2412.16720.

[90] Chris Jenkins, Shantanu Agarwal, Joel Barry, Steven Fincke, and Elizabeth Boschee. Massively multi-lingual event understanding: Extraction, visualization, and search. In Danushka Bollegala, Ruihong Huang, and Alan Ritter, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 247–256, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-demo.23. URL https://aclanthology.org/2023.acl-demo.23.

[91] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

[92] Martin Josifoski, Marija Sakota, Maxime Peyrard, and Robert West. Exploiting asymmetry for synthetic training data generation: SynthIE and the case of information extraction. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1555–1574, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.96. URL https://aclanthology.org/2023.emnlp-main.96.

[93] Thushara Kamalrathne, Dilanthi Amaratunga, Richard Haigh, and Lahiru Kodituwakku. Need for effective detection and early warnings for epidemic and pandemic preparedness planning in the context of multi-hazards: Lessons from the covid-19 pandemic. *International Journal of Disaster Risk Reduction*, 92:103724, 2023.

[94] Emre Kiciman, Robert Osazuwa Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality. *Trans. Mach. Learn. Res.*, 2024, 2024. URL https://openreview.net/forum?id=mqoxLkX210.

[95] Jin-Dong Kim, Yue Wang, Toshihisa Takagi, and Akinori Yonezawa. Overview of Genia event task in BioNLP shared task 2011. In Jun'ichi Tsujii, Jin-Dong Kim, and Sampo Pyysalo, editors, *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 7–15, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL https://aclanthology.org/W11-1802.

[96] Jin-Dong Kim, Yue Wang, and Yamamoto Yasunori. The Genia event extraction shared task, 2013 edition - overview. In Claire Nédellec, Robert Bossy, Jin-Dong Kim, Jung-jae Kim, Tomoko Ohta, Sampo Pyysalo, and Pierre Zweigenbaum, editors, *Proceedings of the BioNLP Shared Task 2013 Workshop*, pages 8–15, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL https://aclanthology.org/W13-2002.

[97] Dorota Kmiec and Frank Kirchhoff. Monkeypox: a new threat? *International journal of molecular sciences*, 23(14):7866, 2022.

[98] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *CoRR*, abs/2205.11916, 2022. doi: 10.48550/ARXIV.2205.11916. URL https://doi.org/10.48550/arXiv.2205.11916.

[99] Xiang Kong, Adithya Renduchintala, James Cross, Yuqing Tang, Jiatao Gu, and Xian Li. Multilingual neural machine translation with deep encoder and multiple shallow decoders. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1613–1624, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.138. URL https://aclanthology.org/2021.eacl-main.138.

[100] Max Kozlov et al. How deadly is monkeypox? what scientists know. *Nature*, 609(7928): 663, 2022.

[101] Jitin Krishnan, Antonios Anastasopoulos, Hemant Purohit, and Huzefa Rangwala. Multilingual code-switching for zero-shot cross-lingual intent prediction and slot filling. In Duygu Ataman, Alexandra Birch, Alexis Conneau, Orhan Firat, Sebastian Ruder, and Gozde Gul Sahin, editors, *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 211–223, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.mrl-1.18. URL https://aclanthology.org/2021.mrl-1.18.

[102] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In Jason Flinn, Margo I. Seltzer, Peter Druschel, Antoine Kaufmann, and Jonathan Mace, editors, *Proceedings of the 29th Symposium on Operating Systems Principles, SOSP 2023, Koblenz, Germany, October 23-26, 2023*, pages 611–626. ACM, 2023. doi: 10.1145/3600006.3613165. URL https://doi.org/10.1145/3600006.3613165.

[103] Alex Lamb, Michael J. Paul, and Mark Dredze. Separating fact from fear: Tracking flu infections on Twitter. In Lucy Vanderwende, Hal Daumé III, and Katrin Kirchhoff, editors, *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 789–795, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL https://aclanthology.org/N13-1097.

[104] Ehsan Latif, Yifan Zhou, Shuchen Guo, Yizhu Gao, Lehong Shi, Matthew Nayaaba, Gyeong-Geon Lee, Liang Zhang, Arne Bewersdorff, Luyang Fang, Xiantong Yang, Huaqin Zhao, Hanqi Jiang, Haoran Lu, Jiaxi Li, Jichao Yu, Weihang You, Zhengliang Liu, Vincent Shung Liu, Hui Wang, Zihao Wu, Jin Lu, Fei Dou, Ping Ma, Ninghao Liu, Tianming Liu, and Xiaoming Zhai. A systematic assessment of openai o1-preview for higher order thinking in education. *CoRR*, abs/2410.21287, 2024. doi: 10.48550/ARXIV.2410.21287. URL https://doi.org/10.48550/arXiv.2410.21287.

[105] Duong Minh Le, Yang Chen, Alan Ritter, and Wei Xu. Constrained decoding for cross-lingual label projection. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=DayPQKXaQk.

[106] Kenton Lee, Kelvin Guu, Luheng He, Tim Dozat, and Hyung Won Chung. Neural data augmentation via example extrapolation. *CoRR*, abs/2102.01335, 2021. URL https://arxiv.org/abs/2102.01335.

[107] Kyungjae Lee, Kyoungho Yoon, Sunghyun Park, and Seung-won Hwang. Semi-supervised training data generation for multilingual question answering. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1437.

[108] Gaël Lejeune, Romain Brixtel, Antoine Doucet, and Nadine Lucas. Multilingual event extraction for epidemic detection. *Artif. Intell. Medicine*, 65(2):131–143, 2015. doi: 10.1016/j.artmed.2015.06.005. URL https://doi.org/10.1016/j.artmed.2015.06.005.

[109] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL https://aclanthology.org/2020.acl-main.703/.

[110] Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. MLQA: Evaluating cross-lingual extractive question answering. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.653. URL https://aclanthology.org/2020.acl-main.653.

[111] Bing Li, Yujie He, and Wenjin Xu. Cross-lingual named entity recognition using parallel corpus: A new approach using xlm-roberta alignment. *CoRR*, abs/2101.11112, 2021. URL https://arxiv.org/abs/2101.11112.

[112] Bo Li, Gexiang Fang, Yang Yang, Quansen Wang, Wei Ye, Wen Zhao, and Shikun Zhang. Evaluating chatgpt's information extraction capabilities: An assessment of performance, explainability, calibration, and faithfulness. *CoRR*, abs/2304.11633, 2023. doi: 10.48550/ARXIV.2304.11633. URL https://doi.org/10.48550/arXiv.2304.11633.

[113] Bo Li, Gexiang Fang, Yang Yang, Quansen Wang, Wei Ye, Wen Zhao, and Shikun Zhang. Evaluating chatgpt's information extraction capabilities: An assessment of performance, explainability, calibration, and faithfulness. *CoRR*, abs/2304.11633, 2023. doi: 10.48550/ARXIV.2304.11633. URL https://doi.org/10.48550/arXiv.2304.11633.

[114] Diya Li, Lifu Huang, Heng Ji, and Jiawei Han. Biomedical event extraction based on knowledge-driven tree-LSTM. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1421–1430, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1145. URL https://aclanthology.org/N19-1145/.

[115] Dongfang Li, Zetian Sun, Xinshuo Hu, Zhenyu Liu, Ziyang Chen, Baotian Hu, Aiguo Wu, and Min Zhang. A survey of large language models attribution. *CoRR*, abs/2311.03731, 2023. doi: 10.48550/ARXIV.2311.03731. URL https://doi.org/10.48550/arXiv.2311.03731.

[116] Fayuan Li, Weihua Peng, Yuguang Chen, Quan Wang, Lu Pan, Yajuan Lyu, and Yong Zhu. Event extraction as multi-turn question answering. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 829–838, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.73. URL https://aclanthology.org/2020.findings-emnlp.73.

[117] Manling Li, Alireza Zareian, Qi Zeng, Spencer Whitehead, Di Lu, Heng Ji, and Shih-Fu Chang. Cross-media structured common space for multimedia event extraction. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2557–2568, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.230. URL https://aclanthology.org/2020.acl-main.230.

[118] Manling Li, Sha Li, Zhenhailong Wang, Lifu Huang, Kyunghyun Cho, Heng Ji, Jiawei Han, and Clare Voss. The future is not one-dimensional: Complex event schema induction by graph modeling for event prediction. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5203–5215, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.422. URL https://aclanthology.org/2021.emnlp-main.422/.

[119] Manling Li, Ruochen Xu, Shuohang Wang, Luowei Zhou, Xudong Lin, Chenguang Zhu, Michael Zeng, Heng Ji, and Shih-Fu Chang. Clip-event: Connecting text and images with event structures. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages

16399–16408. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01593. URL https://doi.org/10.1109/CVPR52688.2022.01593.

[120] Qi Li, Heng Ji, and Liang Huang. Joint event extraction via structured prediction with global features. In Hinrich Schuetze, Pascale Fung, and Massimo Poesio, editors, *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 73–82, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL https://aclanthology.org/P13-1008.

[121] Sha Li, Heng Ji, and Jiawei Han. Document-level event argument extraction by conditional generation. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 894–908, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.69. URL https://aclanthology.org/2021.naacl-main.69.

[122] Sha Li, Qiusi Zhan, Kathryn Conger, Martha Palmer, Heng Ji, and Jiawei Han. GLEN: General-purpose event detection for thousands of types. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2823–2838, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.170. URL https://aclanthology.org/2023.emnlp-main.170.

[123] Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. Synthetic data generation with large language models for text classification: Potential and limitations. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10443–10461, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.647. URL https://aclanthology.org/2023.emnlp-main.647/.

[124] Zixuan Li, Yutao Zeng, Yuxin Zuo, Weicheng Ren, Wenxuan Liu, Miao Su, Yucan Guo, Yantao Liu, Xiang Li, Zhilei Hu, Long Bai, Wei Li, Yidan Liu, Pan Yang, Xiaolong Jin, Jiafeng Guo, and Xueqi Cheng. KnowCoder: Coding structured knowledge into LLMs for universal information extraction. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8758–8779, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.475. URL https://aclanthology.org/2024.acl-long.475/.

[125] Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. A joint neural model for information extraction with global features. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7999–8009, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.713. URL https://aclanthology.org/2020.acl-main.713.

[126] D. A. Lindberg, B. L. Humphreys, and A. T. McCray. The unified medical language system. *Methods of information in medicine*, 32(4):281—-291, 1993. URL https://doi.org/10.1055/s-0038-1634945.

[127] Jian Liu, Yubo Chen, Kang Liu, Wei Bi, and Xiaojiang Liu. Event extraction as machine reading comprehension. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1641–1651, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.128. URL https://aclanthology.org/2020.emnlp-main.128.

[128] Lu Liu and Yifei Fu. Study on the mechanism of public attention to a major event: The outbreak of covid-19 in china. *Sustainable Cities and Society*, 81:103811, 2022.

[129] Keming Lu, I-Hung Hsu, Wenxuan Zhou, Mingyu Derek Ma, and Muhao Chen. Summarization as indirect supervision for relation extraction. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6575–6594, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.490. URL https://aclanthology.org/2022.findings-emnlp.490.

[130] Yaojie Lu, Hongyu Lin, Jin Xu, Xianpei Han, Jialong Tang, Annan Li, Le Sun, Meng Liao, and Shaoyi Chen. Text2Event: Controllable sequence-to-structure generation for end-to-end event extraction. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2795–2806, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.217. URL https://aclanthology.org/2021.acl-long.217/.

[131] Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. Unified structure generation for universal information extraction. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5755–5772, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.395. URL https://aclanthology.org/2022.acl-long.395/.

[132] Kevin Lybarger, Mari Ostendorf, Matthew Thompson, and Meliha Yetisgen. Extracting COVID-19 diagnoses and symptoms from clinical text: A new annotated corpus and neural event extraction framework. *J. Biomed. Informatics*, 117:103761, 2021. doi: 10.1016/j.jbi.2021.103761. URL https://doi.org/10.1016/j.jbi.2021.103761.

[133] Qing Lyu, Hongming Zhang, Elior Sulem, and Dan Roth. Zero-shot event extraction via transfer learning: Challenges and insights. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *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, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.42. URL https://aclanthology.org/2021.acl-short.42.

[134] Mingyu Derek Ma, Xiaoxuan Wang, Po-Nien Kung, P. Jeffrey Brantingham, Nanyun Peng, and Wei Wang. STAR: boosting low-resource information extraction by structure-to-text data generation with large language models. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan, editors, *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 18751–18759. AAAI Press, 2024. doi: 10.1609/AAAI.V38I17.29839. URL https://doi.org/10.1609/aaai.v38i17.29839.

[135] Yubo Ma, Zehao Wang, Yixin Cao, Mukai Li, Meiqi Chen, Kun Wang, and Jing Shao. Prompt for extraction? PAIE: Prompting argument interaction for event argument extraction. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6759–6774, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.466. URL https://aclanthology.org/2022.acl-long.466.

[136] Ryan* Marten, Trung* Vu, Charlie Cheng-Jie Ji, Kartik Sharma, Shreyas Pimpalgaonkar, Alex Dimakis, and Maheswaran Sathiamoorthy. Curator: A Tool for Synthetic Data Creation. https://github.com/bespokelabsai/curator, January 2025.

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

[138] Ethan Mendes, Yang Chen, Wei Xu, and Alan Ritter. Human-in-the-loop evaluation for early misinformation detection: A case study of COVID-19 treatments. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15817–15835, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.881. URL https://aclanthology.org/2023.acl-long.881/.

[139] Zaiqiao Meng, Anya Okhmatovskaia, Maxime Polleri, Yannan Shen, Guido Powell, Zihao Fu, Iris Ganser, Meiru Zhang, Nicholas B. King, David L. Buckeridge, and Nigel Collier. Biocaster in 2021: automatic disease outbreaks detection from global news media. *Bioinform.*, 38(18):4446–4448, 2022. doi: 10.1093/bioinformatics/btac497. URL https://doi.org/10.1093/bioinformatics/btac497.

[140] Bonan Min, Benjamin Rozonoyer, Haoling Qiu, Alexander Zamanian, Nianwen Xue, and Jessica MacBride. ExcavatorCovid: Extracting events and relations from text corpora for temporal and causal analysis for COVID-19. In Heike Adel and Shuming Shi, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language*

Processing: System Demonstrations, pages 63–71, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-demo.8. URL https://aclanthology.org/2021.emnlp-demo.8.

[141] Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. Distant supervision for relation extraction without labeled data. In Keh-Yih Su, Jian Su, Janyce Wiebe, and Haizhou Li, editors, *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011, Suntec, Singapore, August 2009. Association for Computational Linguistics. URL https://aclanthology.org/P09-1113/.

[142] Rishabh Misra. News category dataset. *CoRR*, abs/2209.11429, 2022. doi: 10.48550/ARXIV.2209.11429. URL https://doi.org/10.48550/arXiv.2209.11429.

[143] Mehrad Moradshahi, Giovanni Campagna, Sina Semnani, Silei Xu, and Monica Lam. Localizing open-ontology QA semantic parsers in a day using machine translation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5970–5983, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.481. URL https://aclanthology.org/2020.emnlp-main.481.

[144] Stephen Mutuvi, Antoine Doucet, Gaël Lejeune, and Moses Odeo. A dataset for multi-lingual epidemiological event extraction. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4139–4144, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.509.

[145] Masaaki Nagata, Katsuki Chousa, and Masaaki Nishino. A supervised word alignment method based on cross-language span prediction using multilingual BERT. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 555–565, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.41. URL https://aclanthology.org/2020.emnlp-main.41.

[146] Lukas Netz, Jan Reimer, and Bernhard Rumpe. Using grammar masking to ensure syntactic validity in llm-based modeling tasks. *CoRR*, abs/2407.06146, 2024. doi: 10.48550/ARXIV.2407.06146. URL https://doi.org/10.48550/arXiv.2407.06146.

[147] Thien Huu Nguyen, Kyunghyun Cho, and Ralph Grishman. Joint event extraction via recurrent neural networks. In Kevin Knight, Ani Nenkova, and Owen Rambow, editors, *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, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1034. URL https://aclanthology.org/N16-1034.

[148] Jian Ni and Radu Florian. Neural cross-lingual relation extraction based on bilingual word embedding mapping. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *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 399–409, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1038. URL https://aclanthology.org/D19-1038.

[149] Jian Ni, Georgiana Dinu, and Radu Florian. Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1470–1480, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1135. URL https://aclanthology.org/P17-1135.

[150] Massimo Nicosia, Zhongdi Qu, and Yasemin Altun. Translate & Fill: Improving zero-shot multilingual semantic parsing with synthetic data. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3272–3284, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.279. URL https://aclanthology.org/2021.findings-emnlp.279.

[151] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003. doi: 10.1162/089120103321337421. URL https://aclanthology.org/J03-1002.

[152] Jennifer M Olsen. Epicore: crowdsourcing health professionals to verify disease outbreaks. *Online Journal of Public Health Informatics*, 9(1), 2017.

[153] OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. doi: 10.48550/ARXIV.2303.08774. URL https://doi.org/10.48550/arXiv.2303.08774.

[154] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. *CoRR*, abs/2203.02155, 2022. doi: 10.48550/ARXIV.2203.02155. URL https://doi.org/10.48550/arXiv.2203.02155.

[155] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. *CoRR*, abs/2203.02155, 2022. doi: 10.48550/ARXIV.2203.02155. URL https://doi.org/10.48550/arXiv.2203.02155.

[156] Enny S Paixao, Luciana L Cardim, Maria CN Costa, Elizabeth B Brickley, Rita CO de Carvalho-Sauer, Eduardo H Carmo, Roberto FS Andrade, Moreno S Rodrigues, Rafael V Veiga, Larissa C Costa, et al. Mortality from congenital zika syndrome—nationwide cohort study in brazil. *New England Journal of Medicine*, 386(8): 757–767, 2022.

[157] Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. Unifying large language models and knowledge graphs: A roadmap. *IEEE Trans. Knowl. Data Eng.*, 36(7):3580–3599, 2024. doi: 10.1109/TKDE.2024.3352100. URL https://doi.org/10.1109/TKDE.2024.3352100.

[158] Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. Cross-lingual name tagging and linking for 282 languages. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1178. URL https://aclanthology.org/P17-1178.

[159] Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cícero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. Structured prediction as translation between augmented natural languages. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=US-TP-xnXI.

[160] Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cícero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. Structured prediction as translation between augmented natural languages. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=US-TP-xnXI.

[161] Tanmay Parekh, I-Hung Hsu, Kuan-Hao Huang, Kai-Wei Chang, and Nanyun Peng. GENEVA: Benchmarking generalizability for event argument extraction with hundreds of event types and argument roles. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3664–3686, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.203. URL https://aclanthology.org/2023.acl-long.203.

[162] Tanmay Parekh, I-Hung Hsu, Kuan-Hao Huang, Kai-Wei Chang, and Nanyun Peng. Contextual label projection for structured prediction. In *under review at Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, New Mexico, Mexico, 2024. Association for Computational Linguistics.

[163] Tanmay Parekh, Jeffrey Kwan, Jiarui Yu, Sparsh Johri, Hyosang Ahn, Sreya Muppalla, Kai-Wei Chang, Wei Wang, and Nanyun Peng. SPEED++: A multilingual event extraction framework for epidemic prediction and preparedness. In Yaser Al-Onaizan,

Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12936–12965, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.720. URL https://aclanthology.org/2024.emnlp-main.720/.

[164] Tanmay Parekh, Anh Mac, Jiarui Yu, Yuxuan Dong, Syed Shahriar, Bonnie Liu, Eric Yang, Kuan-Hao Huang, Wei Wang, Nanyun Peng, and Kai-Wei Chang. Event detection from social media for epidemic prediction. In *under review at Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, New Mexico, Mexico, 2024. Association for Computational Linguistics.

[165] Tanmay Parekh, Yuxuan Dong, Lucas Bandarkar, Artin Kim, I-Hung Hsu, Kai-Wei Chang, and Nanyun Peng. SNaRe: Domain-aware data generation for low-resource event detection. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20583–20604, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1039. URL https://aclanthology.org/2025.emnlp-main.1039/.

[166] Tanmay Parekh, Kartik Mehta, Ninareh Mehrabi, Kai-Wei Chang, and Nanyun Peng. DiCoRe: Enhancing zero-shot event detection via divergent-convergent LLM reasoning. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20560–20582, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1038. URL https://aclanthology.org/2025.emnlp-main.1038/.

[167] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. *CoRR*, abs/2304.03442, 2023. doi: 10.48550/ARXIV.2304.03442. URL https://doi.org/10.48550/arXiv.2304.03442.

[168] Kanghee Park, Jiayu Wang, Taylor Berg-Kirkpatrick, Nadia Polikarpova, and Loris D'Antoni. Grammar-aligned decoding. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/2bdc2267c3d7d01523e2e17ac0a754f3-Abstract-Conference.html.

[169] Michael J Paul, Mark Dredze, and David Broniatowski. Twitter improves influenza forecasting. *PLoS currents*, 6, 2014.

[170] Marco Pota, Mirko Ventura, Hamido Fujita, and Massimo Esposito. Multilingual evaluation of pre-processing for bert-based sentiment analysis of tweets. *Expert Systems with Applications*, 181:115119, 2021. ISSN 0957-4174. doi: https://doi.org/10.1016/

j.eswa.2021.115119. URL https://www.sciencedirect.com/science/article/pii/S0957417421005601.

[171] Amir Pouran Ben Veyseh, Javid Ebrahimi, Franck Dernoncourt, and Thien Nguyen. MEE: A novel multilingual event extraction dataset. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9603–9613, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.652. URL https://aclanthology.org/2022.emnlp-main.652.

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

[173] Sampo Pyysalo, Tomoko Ohta, Makoto Miwa, Han-Cheol Cho, Junichi Tsujii, and Sophia Ananiadou. Event extraction across multiple levels of biological organization. *Bioinform.*, 28(18):575–581, 2012. doi: 10.1093/BIOINFORMATICS/BTS407. URL https://doi.org/10.1093/bioinformatics/bts407.

[174] Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. Is ChatGPT a general-purpose natural language processing task solver? In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1339–1384, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.85. URL https://aclanthology.org/2023.emnlp-main.85.

[175] Afshin Rahimi, Yuan Li, and Trevor Cohn. Massively multilingual transfer for NER. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1015. URL https://aclanthology.org/P19-1015.

[176] Alan L Rector, Jeremy E Rogers, and Pam Pole. The galen high level ontology. In *Medical Informatics Europe'96*, pages 174–178. IOS Press, 1996.

[177] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *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 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410.

[178] Christian M Rochefort, David L Buckeridge, and Alan J Forster. Accuracy of using automated methods for detecting adverse events from electronic health record data: a research protocol. *Implementation Science*, 10:1–9, 2015.

[179] Marco Rospocher, Marieke van Erp, Piek Vossen, Antske Fokkens, Itziar Aldabe, German Rigau, Aitor Soroa, Thomas Ploeger, and Tessel Bogaard. Building event-centric knowledge graphs from news. *J. Web Semant.*, 37-38:132–151, 2016. doi: 10.1016/J.WEBSEM.2015.12.004. URL https://doi.org/10.1016/j.websem.2015.12.004.

[180] Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.802. URL https://aclanthology.org/2021.emnlp-main.802.

[181] Sihem Sahnoun and Gaël Lejeune. Multilingual epidemic event extraction : From simple classification methods to open information extraction (OIE) and ontology. In Ruslan Mitkov and Galia Angelova, editors, *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1227–1233, Held Online, September 2021. INCOMA Ltd. URL https://aclanthology.org/2021.ranlp-1.138.

[182] Oscar Sainz, Itziar Gonzalez-Dios, Oier Lopez de Lacalle, Bonan Min, and Eneko Agirre. Textual entailment for event argument extraction: Zero- and few-shot with multi-source learning. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2439–2455, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.187. URL https://aclanthology.org/2022.findings-naacl.187.

[183] Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau, and Eneko Agirre. Gollie: Annotation guidelines improve zero-shot information-extraction. *CoRR*, abs/2310.03668, 2023. doi: 10.48550/ARXIV.2310.03668. URL https://doi.org/10.48550/arXiv.2310.03668.

[184] Taneeya Satyapanich, Francis Ferraro, and Tim Finin. CASIE: extracting cybersecurity event information from text. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8749–8757. AAAI Press, 2020. doi: 10.1609/AAAI.V34I05.6401. URL https://doi.org/10.1609/aaai.v34i05.6401.

[185] Taneeya Satyapanich, Francis Ferraro, and Tim Finin. CASIE: extracting cybersecurity event information from text. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational*

Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, pages 8749–8757. AAAI Press, 2020. doi: 10.1609/AAAI.V34I05.6401. URL https://doi.org/10.1609/aaai.v34i05.6401.

[186] Timo Schick and Hinrich Schütze. Exploiting cloze-questions for few-shot text classification and natural language inference. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.20. URL https://aclanthology.org/2021.eacl-main.20.

[187] Timo Schick and Hinrich Schütze. Generating datasets with pretrained language models. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6943–6951, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.555. URL https://aclanthology.org/2021.emnlp-main.555/.

[188] Timo Schick and Hinrich Schütze. It's not just size that matters: Small language models are also few-shot learners. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.185. URL https://aclanthology.org/2021.naacl-main.185.

[189] Lynn M. Schriml, James B. Munro, Mike Schor, Dustin Olley, Carrie McCracken, Victor Felix, J. Allen Baron, Rebecca C. Jackson, Susan M. Bello, Cynthia Bearer, Richard Lichenstein, Katharine Bisordi, Nicole Campion, Michelle G. Giglio, and Carol Greene. The human disease ontology 2022 update. *Nucleic Acids Res.*, 50(D1):1255–1261, 2022. doi: 10.1093/nar/gkab1063. URL https://doi.org/10.1093/nar/gkab1063.

[190] Yunfan Shao, Linyang Li, Yichuan Ma, Peiji Li, Demin Song, Qinyuan Cheng, Shimin Li, Xiaonan Li, Pengyu Wang, Qipeng Guo, Hang Yan, Xipeng Qiu, Xuanjing Huang, and Dahua Lin. Case2Code: Scalable synthetic data for code generation. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11056–11069, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.733/.

[191] Jiaming Shen, Yunyi Zhang, Heng Ji, and Jiawei Han. Corpus-based open-domain event type induction. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5427–5440, Online and Punta Cana, Dominican

Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.441. URL https://aclanthology.org/2021.emnlp-main.441.

[192] Tom Sherborne and Mirella Lapata. Zero-shot cross-lingual semantic parsing. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4134–4153, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.285. URL https://aclanthology.org/2022.acl-long.285.

[193] Fatemeh Shiri, Farhad Moghimifar, Reza Haffari, Yuan-Fang Li, Van Nguyen, and John Yoo. Decompose, enrich, and extract! schema-aware event extraction using llms. In *27th International Conference on Information Fusion, FUSION 2024, Venice, Italy, July 8-11, 2024*, pages 1–8. IEEE, 2024. doi: 10.23919/FUSION59988.2024.10706385. URL https://doi.org/10.23919/FUSION59988.2024.10706385.

[194] Alessio Signorini, Alberto Maria Segre, and Philip M Polgreen. The use of twitter to track levels of disease activity and public concern in the us during the influenza a h1n1 pandemic. *PloS one*, 6(5):e19467, 2011.

[195] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. *CoRR*, abs/2408.03314, 2024. doi: 10.48550/ARXIV.2408.03314. URL https://doi.org/10.48550/arXiv.2408.03314.

[196] Zhiyi Song, Ann Bies, Stephanie Strassel, Tom Riese, Justin Mott, Joe Ellis, Jonathan Wright, Seth Kulick, Neville Ryant, and Xiaoyi Ma. From light to rich ERE: Annotation of entities, relations, and events. In Eduard Hovy, Teruko Mitamura, and Martha Palmer, editors, *Proceedings of the The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 89–98, Denver, Colorado, June 2015. Association for Computational Linguistics. doi: 10.3115/v1/W15-0812. URL https://aclanthology.org/W15-0812.

[197] Zirui Song, Bin Yan, Yuhan Liu, Miao Fang, Mingzhe Li, Rui Yan, and Xiuying Chen. Injecting domain-specific knowledge into large language models: A comprehensive survey. *CoRR*, abs/2502.10708, 2025. doi: 10.48550/ARXIV.2502.10708. URL https://doi.org/10.48550/arXiv.2502.10708.

[198] Saurabh Srivastava, Sweta Pati, and Ziyu Yao. Instruction-tuning llms for event extraction with annotation guidelines. *CoRR*, abs/2502.16377, 2025. doi: 10.48550/ARXIV.2502.16377. URL https://doi.org/10.48550/arXiv.2502.16377.

[199] Michael Q. Stearns, Colin Price, Kent A. Spackman, and Amy Y. Wang. SNOMED clinical terms: overview of the development process and project status. In *AMIA 2001, American Medical Informatics Association Annual Symposium, Washington, DC, USA, November 3-7, 2001*. AMIA, 2001. URL https://knowledge.amia.org/amia-55142-a2001a-1.597057/t-001-1.599654/f-001-1.599655/a-133-1.599740/a-134-1.599737.

[200] Elias Stengel-Eskin, Tzu-ray Su, Matt Post, and Benjamin Van Durme. A discriminative neural model for cross-lingual word alignment. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *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 910–920, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1084. URL https://aclanthology.org/D19-1084.

[201] Ananya Subburathinam, Di Lu, Heng Ji, Jonathan May, Shih-Fu Chang, Avirup Sil, and Clare Voss. Cross-lingual structure transfer for relation and event extraction. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *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 313–325, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1030. URL https://aclanthology.org/D19-1030.

[202] Jiao Sun and Nanyun Peng. Men are elected, women are married: Events gender bias on Wikipedia. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *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 350–360, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.45. URL https://aclanthology.org/2021.acl-short.45.

[203] Zhaoyue Sun, Jiazheng Li, Gabriele Pergola, Byron Wallace, Bino John, Nigel Greene, Joseph Kim, and Yulan He. PHEE: A dataset for pharmacovigilance event extraction from text. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5571–5587, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.376. URL https://aclanthology.org/2022.emnlp-main.376.

[204] Beth M. Sundheim. Overview of the fourth Message Understanding Evaluation and Conference. In *Fourth Message Understanding Conference (MUC-4): Proceedings of a Conference Held in McLean, Virginia, June 16-18, 1992*, 1992. URL https://aclanthology.org/M92-1001.

[205] Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung-yi Lee, and Yun-Nung Chen. Let me speak freely? a study on the impact of format restrictions on large language model performance. In Franck Dernoncourt, Daniel Preoţiuc-Pietro, and Anastasia Shimorina, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1218–1236, Miami, Florida, US, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-industry.91. URL https://aclanthology.org/2024.emnlp-industry.91/.

[206] Ruixiang Tang, Xiaotian Han, Xiaoqian Jiang, and Xia Hu. Does synthetic data

generation of llms help clinical text mining? *CoRR*, abs/2303.04360, 2023. doi: 10.48550/ARXIV.2303.04360. URL https://doi.org/10.48550/arXiv.2303.04360.

[207] Simone Tedeschi and Roberto Navigli. MultiNERD: A multilingual, multi-genre and fine-grained dataset for named entity recognition (and disambiguation). In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 801–812, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.60. URL https://aclanthology.org/2022.findings-naacl.60.

[208] Nirmalya Thakur. Monkeypox2022tweets: A large-scale twitter dataset on the 2022 monkeypox outbreak, findings from analysis of tweets, and open research questions. *Infectious Disease Reports*, 14(6):855–883, 2022. URL https://pubmed.ncbi.nlm.nih.gov/36412745/.

[209] Yufei Tian, Abhilasha Ravichander, Lianhui Qin, Ronan Le Bras, Raja Marjieh, Nanyun Peng, Yejin Choi, Thomas Griffiths, and Faeze Brahman. MacGyver: Are large language models creative problem solvers? In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5303–5324, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.297. URL https://aclanthology.org/2024.naacl-long.297/.

[210] Erik F. Tjong Kim Sang. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*, 2002. URL https://aclanthology.org/W02-2024.

[211] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147, 2003. URL https://aclanthology.org/W03-0419.

[212] MeiHan Tong, Bin Xu, Shuai Wang, Meihuan Han, Yixin Cao, Jiangqi Zhu, Siyu Chen, Lei Hou, and Juanzi Li. DocEE: A large-scale and fine-grained benchmark for document-level event extraction. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *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, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.291. URL https://aclanthology.org/2022.naacl-main.291.

[213] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971, 2023. doi: 10.48550/ARXIV.2302.13971. URL https://doi.org/10.48550/arXiv.2302.13971.

[214] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[215] Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. Context-aware neural machine translation learns anaphora resolution. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1264–1274, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1117. URL https://aclanthology.org/P18-1117.

[216] David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. Entity, relation, and event extraction with contextualized span representations. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *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, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1585. URL https://aclanthology.org/D19-1585.

[217] Qing Wang, Kang Zhou, Qiao Qiao, Yuepei Li, and Qi Li. Improving unsupervised relation extraction by augmenting diverse sentence pairs. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12136–12147, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.745. URL https://aclanthology.org/2023.emnlp-main.745/.

[218] Sijia Wang, Mo Yu, Shiyu Chang, Lichao Sun, and Lifu Huang. Query and extract: Refining event extraction as type-oriented binary decoding. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 169–182, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.16. URL https://aclanthology.org/2022.findings-acl.16.

[219] Xiaoyan Wang, George Hripcsak, Marianthi Markatou, and Carol Friedman. Research paper: Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: A feasibility study. *J. Am. Medical Informatics Assoc.*, 16(3):328–337, 2009. doi: 10.1197/JAMIA.M3028. URL https://doi.org/10.1197/jamia.M3028.

[220] Xiaozhi Wang, Ziqi Wang, Xu Han, Wangyi Jiang, Rong Han, Zhiyuan Liu, Juanzi Li, Peng Li, Yankai Lin, and Jie Zhou. MAVEN: A Massive General Domain Event Detection Dataset. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1652–1671, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.129. URL https://aclanthology.org/2020.emnlp-main.129.

[221] Xingyao Wang, Sha Li, and Heng Ji. Code4Struct: Code generation for few-shot event structure prediction. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3640–3663, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.202. URL https://aclanthology.org/2023.acl-long.202.

[222] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.754. URL https://aclanthology.org/2023.acl-long.754/.

[223] Zirui Wang, Adams Wei Yu, Orhan Firat, and Yuan Cao. Towards zero-label language learning. *CoRR*, abs/2109.09193, 2021. URL https://arxiv.org/abs/2109.09193.

[224] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903, 2022. URL https://arxiv.org/abs/2201.11903.

[225] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

[226] Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, Yong Jiang, and Wenjuan Han. Zero-shot information extraction via chatting with chatgpt. *CoRR*, abs/2302.10205, 2023. doi: 10.48550/ARXIV.2302.10205. URL https://doi.org/10.48550/arXiv.2302.10205.

[227] Brandon T. Willard and Rémi Louf. Efficient guided generation for large language models. *CoRR*, abs/2307.09702, 2023. doi: 10.48550/ARXIV.2307.09702. URL https://doi.org/10.48550/arXiv.2307.09702.

[228] KayYen Wong, Sameen Maruf, and Gholamreza Haffari. Contextual neural machine translation improves translation of cataphoric pronouns. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5971–5978, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.530. URL https://aclanthology.org/2020.acl-main.530.

[229] Likang Wu, Zhi Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, Hui Xiong, and Enhong Chen. A survey on large language models for recommendation. *World Wide Web (WWW)*, 27(5):60, 2024. doi: 10.1007/S11280-024-01291-2. URL https://doi.org/10.1007/s11280-024-01291-2.

[230] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. Autogen: Enabling next-gen LLM applications via multi-agent conversation framework. *CoRR*, abs/2308.08155, 2023. doi: 10.48550/ARXIV.2308.08155. URL https://doi.org/10.48550/arXiv.2308.08155.

[231] Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, and Enhong Chen. Large language models for generative information extraction: A survey. *CoRR*, abs/2312.17617, 2023. doi: 10.48550/ARXIV.2312.17617. URL https://doi.org/10.48550/arXiv.2312.17617.

[232] Weijia Xu, Batool Haider, and Saab Mansour. End-to-end slot alignment and recognition for cross-lingual NLU. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5052–5063, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.410. URL https://aclanthology.org/2020.emnlp-main.410.

[233] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.41. URL https://aclanthology.org/2021.naacl-main.41.

[234] Kabir Yadav, Efsun Sarioglu, Meaghan Smith, and Hyeong-Ah Choi. Automated outcome classification of emergency department computed tomography imaging reports. *Academic Emergency Medicine*, 20(8):848–854, 2013.

[235] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *CoRR*, abs/2412.15115, 2024. doi: 10.48550/ARXIV.2412.15115. URL https://doi.org/10.48550/arXiv.2412.15115.

[236] Bishan Yang and Tom M. Mitchell. Joint extraction of events and entities within a document context. In Kevin Knight, Ani Nenkova, and Owen Rambow, editors,

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, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1033. URL https://aclanthology.org/N16-1033.

[237] Jian Yang, Shaohan Huang, Shuming Ma, Yuwei Yin, Li Dong, Dongdong Zhang, Hongcheng Guo, Zhoujun Li, and Furu Wei. CROP: Zero-shot cross-lingual named entity recognition with multilingual labeled sequence translation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 486–496, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.34. URL https://aclanthology.org/2022.findings-emnlp.34.

[238] Sen Yang, Dawei Feng, Linbo Qiao, Zhigang Kan, and Dongsheng Li. Exploring pre-trained language models for event extraction and generation. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5284–5294, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1522. URL https://aclanthology.org/P19-1522.

[239] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *CoRR*, abs/2210.03629, 2022. doi: 10.48550/ARXIV.2210.03629. URL https://doi.org/10.48550/arXiv.2210.03629.

[240] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/271db9922b8d1f4dd7aaef84ed5ac703-Abstract-Conference.html.

[241] Wenlin Yao, Xiaoman Pan, Lifeng Jin, Jianshu Chen, Dian Yu, and Dong Yu. Connect-the-dots: Bridging semantics between words and definitions via aligning word sense inventories. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7741–7751, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.610. URL https://aclanthology.org/2021.emnlp-main.610.

[242] Mahsa Yarmohammadi, Shijie Wu, Marc Marone, Haoran Xu, Seth Ebner, Guanghui Qin, Yunmo Chen, Jialiang Guo, Craig Harman, Kenton Murray, Aaron Steven White, Mark Dredze, and Benjamin Van Durme. Everything is all it takes: A multipronged strategy for zero-shot cross-lingual information extraction. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the*

2021 Conference on Empirical Methods in Natural Language Processing, pages 1950–1967, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.149. URL https://aclanthology.org/2021.emnlp-main.149.

[243] David Yarowsky, Grace Ngai, and Richard Wicentowski. Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the First International Conference on Human Language Technology Research*, 2001. URL https://aclanthology.org/H01-1035.

[244] Alexander Yates, Michele Banko, Matthew Broadhead, Michael Cafarella, Oren Etzioni, and Stephen Soderland. TextRunner: Open information extraction on the web. In Bob Carpenter, Amanda Stent, and Jason D. Williams, editors, *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 25–26, Rochester, New York, USA, April 2007. Association for Computational Linguistics. URL https://aclanthology.org/N07-4013/.

[245] Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. ZeroGen: Efficient zero-shot learning via dataset generation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11653–11669, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.801. URL https://aclanthology.org/2022.emnlp-main.801/.

[246] Pengfei Yu, Jonathan May, and Heng Ji. Bridging the gap between native text and translated text through adversarial learning: A case study on cross-lingual event extraction. In Andreas Vlachos and Isabelle Augenstein, editors, *Findings of the Association for Computational Linguistics: EACL 2023*, pages 754–769, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-eacl.57. URL https://aclanthology.org/2023.findings-eacl.57.

[247] Rodolfo Zevallos, John Ortega, William Chen, Richard Castro, Núria Bel, Cesar Toshio, Renzo Venturas, Hilario Aradiel, and Nelsi Melgarejo. Introducing QuBERT: A large monolingual corpus and BERT model for Southern Quechua. In Colin Cherry, Angela Fan, George Foster, Gholamreza (Reza) Haffari, Shahram Khadivi, Nanyun (Violet) Peng, Xiang Ren, Ehsan Shareghi, and Swabha Swayamdipta, editors, *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 1–13, Hybrid, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.deeplo-1.1. URL https://aclanthology.org/2022.deeplo-1.1.

[248] Honghua Zhang, Po-Nien Kung, Masahiro Yoshida, Guy Van den Broeck, and Nanyun Peng. Adaptable logical control for large language models. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Pro-*

cessing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/d15c16cf5619a2b1606da5fc88e3f1a9-Abstract-Conference.html.

[249] Hongming Zhang, Xin Liu, Haojie Pan, Yangqiu Song, and Cane Wing-Ki Leung. ASER: A large-scale eventuality knowledge graph. In Yennun Huang, Irwin King, Tie-Yan Liu, and Maarten van Steen, editors, *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, pages 201–211. ACM / IW3C2, 2020. doi: 10.1145/3366423.3380107. URL https://doi.org/10.1145/3366423.3380107.

[250] Hongming Zhang, Haoyu Wang, and Dan Roth. Zero-shot Label-aware Event Trigger and Argument Classification. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1331–1340, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.114. URL https://aclanthology.org/2021.findings-acl.114.

[251] Hongming Zhang, Wenlin Yao, and Dong Yu. Efficient zero-shot event extraction with context-definition alignment. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7169–7179, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.531. URL https://aclanthology.org/2022.findings-emnlp.531.

[252] Weiyan Zhang, Wanpeng Lu, Jiacheng Wang, Yating Wang, Lihan Chen, Haiyun Jiang, Jingping Liu, and Tong Ruan. Unexpected phenomenon: LLMs' spurious associations in information extraction. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9176–9190, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.545. URL https://aclanthology.org/2024.findings-acl.545/.

[253] Zixuan Zhang and Heng Ji. Abstract Meaning Representation guided graph encoding and decoding for joint information extraction. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 39–49, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.4. URL https://aclanthology.org/2021.naacl-main.4.

[254] Zhihan Zhou, Liqian Ma, and Han Liu. Trade the event: Corporate events detection for news-based event-driven trading. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2114–2124, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.186. URL https://aclanthology.org/2021.findings-acl.186/.

[255] Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Jiajun Chen, Lei Li, and Shujian Huang. Multilingual machine translation with large language models: Empirical results and analysis. *CoRR*, abs/2304.04675, 2023. doi: 10.48550/arXiv.2304.04675. URL https://doi.org/10.48550/arXiv.2304.04675.

[256] Mingchen Zhuge, Haozhe Liu, Francesco Faccio, Dylan R. Ashley, Róbert Csordás, Anand Gopalakrishnan, Abdullah Hamdi, Hasan Abed Al Kader Hammoud, Vincent Herrmann, Kazuki Irie, Louis Kirsch, Bing Li, Guohao Li, Shuming Liu, Jinjie Mai, Piotr Piekos, Aditya A. Ramesh, Imanol Schlag, Weimin Shi, Aleksandar Stanic, Wenyi Wang, Yuhui Wang, Mengmeng Xu, Deng-Ping Fan, Bernard Ghanem, and Jürgen Schmidhuber. Mindstorms in natural language-based societies of mind. *Comput. Vis. Media*, 11(1):29–81, 2025. doi: 10.26599/CVM.2025.9450460. URL https://doi.org/10.26599/cvm.2025.9450460.

[257] Shi Zong, Ashutosh Baheti, Wei Xu, and Alan Ritter. Extracting a knowledge base of COVID-19 events from social media. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3810–3823, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.335.

[258] Jinan Zou, Yanxi Liu, Yuankai Qi, Haiyao Cao, Lingqiao Liu, and Javen Qinfeng Shi. A generative approach for comprehensive financial event extraction at the document level. In *4th ACM International Conference on AI in Finance, ICAIF 2023, Brooklyn, NY, USA, November 27-29, 2023*, pages 323–330. ACM, 2023. doi: 10.1145/3604237.3626844. URL https://doi.org/10.1145/3604237.3626844.