New Frontiers of Information Extraction

Muhao Chen¹, Lifu Huang², Manling Li³, Ben Zhou⁴, Heng Ji³,⁵, Dan Roth⁴,⁶
¹Department of Computer Science, USC
²Department of Computer Science, Virginia Tech
³Department of Computer Science, UIUC
⁴Department of Computer and Information Science, UPenn
⁵Alexa AI & ⁶AWS AI Labs

Abstract
This tutorial targets researchers and practitioners who are interested in AI and ML technologies for structural information extraction (IE) from unstructured textual sources. In particular, this tutorial will provide audience with a systematic introduction to recent advances in IE, by addressing several important research questions. These questions include (i) how to develop a robust IE system from a small amount of noisy training data, while ensuring the reliability of its prediction? (ii) how to foster the generalizability of IE through enhancing the system’s cross-lingual, cross-domain, cross-task and cross-modal transferability? (iii) how to support extracting structural information with extremely fine-grained and diverse labels? (iv) how to further improve IE by leveraging indirect supervision from other NLP tasks, such as Natural Language Generation (NLG), Natural Language Inference (NLI), Question Answering (QA) or summarization, and pre-trained language models? (v) how to acquire knowledge to guide inference in IE systems? We will discuss several lines of frontier research that tackle those challenges, and will conclude the tutorial by outlining directions for further investigation.

1 Introduction
Information extraction (IE) is the process of automatically extracting structural information from unstructured or semi-structured data. It provides the essential support for natural language understanding by recognizing and resolving the concepts, entities, events described in text, and inferring the relations among them. In various application domains, IE automates the costly acquisition process of domain-specific knowledge representations that have been the backbone of any knowledge-driven AI systems. For example, automated knowledge base construction has relied on technologies for entity-centric IE (Carlson et al., 2010; Lehmann et al., 2015). Extraction of events and event chains assists machines with narrative prediction (Zhang et al., 2021b; Chaturvedi et al., 2017) and summarization tasks (Liu et al., 2018; Chen et al., 2019b). Medical IE also benefits important but expensive clinical tasks such as drug discovery and repurposing (Sosa et al., 2019; Munkhdalai et al., 2018). Despite the importance, frontier research in IE still faces several key challenges. The first challenge is that existing dominant methods using language modeling representation cannot sufficiently capture the essential knowledge and structures required for IE tasks. The second challenge is on the development of extraction models for fine-grained information with less supervision, considering that obtaining structural annotation on unlabeled data has been very costly. The third challenge is to extend the reliability and generalizability of IE systems in real-world scenarios, where data sources often contain incorrect, invalid or unrecognizable inputs, as well as inputs containing unseen labels and mixture of modalities. By tackling those critical challenges, recent literature is leading to transformative advancement in principles and methodologies of IE system development. We believe it is necessary to present a timely tutorial to comprehensively summarize the new frontiers in IE research and point out the emerging challenges that deserve further investigation.

In this tutorial, we will systematically review several lines of frontier research on developing robust, reliable and adaptive learning systems for extracting rich structured information. Beyond introducing robust learning and inference methods for unsupervised denoising, constraint capture and novelty detection, we will discuss recent approaches for leveraging indirect supervision from natural language inference and generation tasks to improve IE. We will also review recent minimally supervised methods for training IE models with distant supervision from linguistic patterns, corpus statistics or language modeling objectives. In addition, we will
illustrate how a model trained on a close domain can be reliably adapted to produce extraction from data sources in different domains, languages and modalities, or acquiring global knowledge (e.g., event schemas) to guide the extraction on a highly diverse open label space. Participants will learn about recent trends and emerging challenges in this topic, representative tools and learning resources to obtain ready-to-use models, and how related technologies benefit end-user NLP applications. A graphical abstract of this tutorial is provided as Fig. 1, which serves as our roadmap for new frontiers of information extraction.

2 Outline of Tutorial Content

This half-day tutorial presents a systematic overview of recent advancement in IE technologies. We will begin motivating this topic with a selection of real-world applications and emerging challenges of IE. Then, we will introduce robust learning methods and inference methods to tackle noisy supervision, prediction inconsistency and out-of-distribution (OOD) inputs. We will also discuss about indirect supervision and minimal supervision methods that further improves IE model development under limited learning resources. Based on the robust IE systems developed in close-domain settings, we will explain how transfer learning technologies can adaptively extend the utility of the systems across domains, languages and tasks, and how complementary information can be extracted from data modalities other than human language. Moreover, we will exemplify the use of aforementioned technologies in various end-user NLP applications such as misinformation detection and scientific discovery, and will outline emerging research challenges that may catalyze further investigation on developing reliable and adaptive learning systems for IE. The detailed contents are outlined below.

2.1 Background and Motivation [20min]

We will define the main research problem and motivate the topic by presenting several real-world NLP and knowledge-driven AI applications of IE technologies, as well as several key challenges that are at the core of frontier research in this area.

2.2 Robust Learning and Inference for IE [35min]

We will introduce methodologies that enhance the robustness of learning systems for IE in both their learning and inference phases. Those methodologies involve self-supervised denoising techniques for training noise-robust IE models based on co-regularized knowledge distillation (Zhou and Chen, 2021; Liang et al., 2021), label re-weighting (Wang et al., 2019b) and label smoothing (Lukasik et al., 2020). Besides, we will also discuss about unsuper-
vised techniques for out-of-distribution (OOD) detection (Zhou et al., 2021b; Hendrycks et al., 2020), prediction with abstention (Dhamija et al., 2018; Hendrycks et al., 2018) and novelty class detection (Perera and Patel, 2019) that seek to help the IE model identify invalid inputs or inputs with semantic shifts during its inference phase. Specifically, to demonstrate how models can ensure the global consistency of the extraction, we will cover constraint learning methods that automatically capture logical constraints among relations (Wang et al., 2021a, 2022c; Pan et al., 2020), and techniques to enforce the constraints in inference (Wang et al., 2020; Li et al., 2019a; Han et al., 2019; Lin et al., 2020). To assess if the systems give faithful extracts, we will also talk about the spurious correlation problems of current IE models and how to address them with counterfactual analysis (Wang et al., 2022b; Qian et al., 2021).

2.3 Minimally and Indirectly Supervised IE [35min]

We will introduce effective approaches that use alternative supervision sources for IE, that is, to use supervision signals from related tasks to make up for the lack of quantity and comprehensiveness in IE-specific training data. This includes indirect supervision sources such as question answering and reading comprehension (Wu et al., 2020; Lyu et al., 2021; Levy et al., 2017; Li et al., 2019b; Du and Cardie, 2020), natural language inference (Li et al., 2022a; Yin et al., 2020) and generation (Lu et al., 2021; Li et al., 2021b). We will also cover the use of weak supervision sources such as structural texts (e.g., Wikipedia) (Ji et al., 2017; Zhou et al., 2018) and global biases (Ning et al., 2018b). With the breakthrough of large-scale pre-trained language models (Devlin et al., 2019; Li et al., 2022c), methodologies have been proposed to explore the language model objective as indirect supervision for IE. To this end, we will cover methods includes direct probing (Feldman et al., 2019; Zhang et al., 2020c), and more recently, pre-training with distant signals acquired from linguistic patterns (Zhou et al., 2020, 2021a).

2.4 Transferability of IE Systems [35min]

One important challenge of developing IE systems lies in the limited coverage of predefined schemas (e.g., predefined types of entities, relations or events) and the heavy reliance on human annotations. When moving to new types, domains or languages, we have to start from scratch by creating annotations and re-training the extraction models. In this part of tutorial, we will cover the recent advances in improving the transferability of IE, including (1) cross-lingual transfer by leveraging adversarial training (Chen et al., 2019a; Huang et al., 2019; Zhou et al., 2019), language-invariant representations (Huang et al., 2018a; Subburathinam et al., 2019) and resources (Tsai et al., 2016; Pan et al., 2017), pre-trained multilingual language models (Wu and Dredze, 2019; Conneau et al., 2020) as well as data projection (Ni et al., 2017; Yarmohammadi et al., 2021), (2) cross-type transfer including zero-shot and few-shot IE by learning prototypes (Huang et al., 2018b; Chan et al., 2019; Huang and Ji, 2020), reading the definitions (Chen et al., 2020b; Logeswaran et al., 2019; Obeidat et al., 2019; Yu et al., 2022; Wang et al., 2022a), answering questions (Levy et al., 2017; Liu et al., 2020; Lyu et al., 2021), and (3) transfer across different benchmark datasets (Xia and Van Durme, 2021; Wang et al., 2021b). Finally, we will also discuss the progress on life-long learning for IE (Wang et al., 2019a; Cao et al., 2020; Yu et al., 2021; Liu et al., 2022) to enable knowledge transfer across incrementally updated models.

2.5 Cross-modal IE [20min]

Cross-modal IE aims to extract structured knowledge from multiple modalities, including unstructured and semi-structured text, images, videos, tables, etc. We will start from visual event and argument extraction from images (Yatskar et al., 2016; Gkioxari et al., 2018; Pratt et al., 2020; Zareian et al., 2020; Li et al., 2022b) and videos (Gu et al., 2018; Sadhu et al., 2021; Chen et al., 2021a). To extract multimedia events, the key challenge is to identify the cross-modal coreference and linking (Deng et al., 2018; Akbari et al., 2018; Pratt et al., 2020; Zareian et al., 2020; Li et al., 2022b) and videos (Gu et al., 2018; Sadhu et al., 2021; Chen et al., 2021a). To extract multimedia events, the key challenge is to identify the cross-modal coreference and linking (Deng et al., 2018; Akbari et al., 2018; Zeng et al., 2019) and represent both text and visual knowledge in a common semantic space (Li et al., 2020a; Chen et al., 2021a). To extract multimedia events, the key challenge is to identify the cross-modal coreference and linking (Deng et al., 2018; Akbari et al., 2018; Zeng et al., 2019) and represent both text and visual knowledge in a common semantic space (Li et al., 2020a; Chen et al., 2021a; Zhang et al., 2021a; Li et al., 2022b). We will also introduce the information extraction from semi-structured data (Katti et al., 2018; Qian et al., 2019) and tabular data (Herzig et al., 2020).

2.6 Knowledge-guided IE [15min]

Global knowledge representation induced from large-scale corpora can guide the inference about the complicated connections between knowledge elements and help fix the extraction errors. We will
introduce cross-task and cross-instance statistical constraint knowledge (Lin et al., 2020; Van Nguyen et al., 2021), commonsense knowledge (Ning et al., 2018a), and global event schema knowledge (Li et al., 2020b; Wen et al., 2021; Li et al., 2021a; Jin et al., 2022) that help jointly extract entities, relations, and events.

2.7 Future Research Directions [30min]
IE is a key component in supporting knowledge acquisition and it impacts a wide spectrum of knowledge-driven AI applications. We will conclude the tutorial by presenting further challenges and potential research topics in identifying trustworthiness of extracted content (Zhang et al., 2019, 2020b), IE with quantitative reasoning (Elazar et al., 2019; Zhang et al., 2020a), cross-document IE (Caciularu et al., 2021), incorporating domain-specific knowledge (Lai et al., 2021; Zhang et al., 2021c), extension to knowledge reasoning and prediction, modeling of label semantics (Huang et al., 2022; Mueller et al., 2022; Ma et al., 2022; Chen et al., 2020a), and challenges for acquiring implicit but essential information from corpora that potentially involve reporting bias (Sap et al., 2020).

3 Specification of the Tutorial
The proposed tutorial is considered a cutting-edge tutorial that introduces new frontiers in IE research. The presented topic has not been covered by ACL/EMNLP/NAACL/EACL/COLING tutorials in the past 4 years. One exception is the ACL 2020 tutorial “Multi-modal Information Extraction from Text, Semi-structured, and Tabular Data on the Web” that is partly relevant to one of our technical sections (§2.5). That particular section of our talk will focus on IE from visual and multi-media data in addition to semi-structured data, being different from the aforementioned ACL 2020 tutorial that has mainly covered topics on semi-structured data.

Audience and Prerequisites Based on the level of interest in this topic, we expect around 150 participants. While no specific background knowledge is assumed of the audience, it would be the best for the attendees to know about basic deep learning technologies, pre-trained word embeddings (e.g. Word2Vec) and language models (e.g. BERT). A reading list that could help provide background knowledge to the audience before attending this tutorial is given in Appx. §A.2.

4 Tutorial Instructors
The following are biographies of the speaker. Past tutorials given by us are listed in Appx. §A.1.

Muha Chen is an Assistant Research Professor of Computer Science at USC, where he directs the Language Understanding and Knowledge Acquisition (LUKA) Group. His research focuses on data-driven machine learning approaches for natural language understanding and knowledge acquisition. His work has been recognized with an NSF CRII Award, a Cisco Faculty Research Award, an ACM SIGBio Best Student Paper Award, and a Best Paper Nomination at CoNLL. Muha obtained his B.S. in Computer Science degree from Fudan University in 2014, his PhD degree from UCLA Department of Computer Science in 2019, and was a postdoctoral researcher at UPenn prior to joining USC. Additional information is available at http://muhaochen.github.io.

Lifu Huang is an Assistant Professor at the Computer Science department of Virginia Tech. He obtained a PhD in Computer Science from UIUC. He has a wide range of research interests in NLP, including extracting structured knowledge with limited supervision, natural language understanding and reasoning with external knowledge and commonsense, natural language generation, representation learning for cross-lingual and cross-domain transfer, and multi-modality learning. He is a recipient of the 2019 AI2 Fellowship and 2021 Amazon Research Award. Additional information is available at https://wilburone.github.io.

Manling Li is a fourth-year Ph.D. student at the Computer Science Department of UIUC. Manling has won the Best Demo Paper Award at ACL’20, the Best Demo Paper Award at NAACL’21, C.L. Dave and Jane W.S. Liu Award, and has been selected as Mavis Future Faculty Fellow. She is a recipient of Microsoft Research PhD Fellowship. She has more than 30 publications on knowledge extraction and reasoning from multi-media data. Additional information is available at https://limanling.github.io.

Ben Zhou is a third-year Ph.D. student at the Department of Computer and Information Science, University of Pennsylvania. He obtained his B.S. from UIUC in 2019. Ben’s research interests...
are distant supervision extraction and experiential knowledge reasoning, and he has more than 5 recent papers on related topics. He is a recipient of the ENIAC fellowship from the University of Pennsylvania, and a finalist of the CRA outstanding undergraduate researcher award. Additional information is available at http://xuanyu.me/.

**Heng Ji** is a Professor at Computer Science Department of University of Illinois Urbana-Champaign, and an Amazon Scholar. She received her B.A. and M. A. in Computational Linguistics from Tsinghua University, and her M.S. and Ph.D. in Computer Science from New York University. Her research interests focus on NLP, especially on Multimedia Multilingual Information Extraction, Knowledge Base Population and Knowledge-driven Generation. She was selected as “Young Scientist” and a member of the Global Future Council on the Future of Computing by the World Economic Forum. The awards she received include “AI’s 10 to Watch” Award, NSF CAREER award, Google Research Award, IBM Watson Faculty Award, Bosch Research Award, and Amazon AWS Award, ACL2020 Best Demo Paper Award, and NAACL2021 Best Demo Paper Award. Additional information is available at https://blender.cs.illinois.edu/hengji.html.

**Dan Roth** is the Eduardo D. Glandt Distinguished Professor at the Department of Computer and Information Science, UPenn, the NLP Lead at AWS AI Labs, and a Fellow of the AAAS, ACM, AAAI, and ACL. In 2017 Roth was awarded the John McCarthy Award, the highest award the AI community gives to mid-career AI researchers. Roth was recognized “for major conceptual and theoretical advances in the modeling of natural language understanding, machine learning, and reasoning.” Roth has published broadly in NLP, KRR, and learning theory, and has given keynote talks and tutorials in all ACL and AAAI major conferences. Roth was the Editor-in-Chief of JAIR until 2017, and was the program chair of AAAI’11, ACL’03 and CoNLL’02; he serves regularly as an area chair and senior program committee member in the major conferences in his research areas. Prof. Roth received his B.A Summa cum laude in Mathematics from the Technion, and his Ph.D. in Computer Science from Harvard University in 1995. Additional information is available at http://www.cis.upenn.edu/~danroth/.

**Acknowledgement**

This presenters’ research is supported in part by U.S. DARPA KAIROS Program No. FA8750-19-2-1004, DARPA AIDA Program No. FA8750-18-2-0014, DARPA MCS program No. N660011924033, and by the NSF of United States Grant IIS 2105329. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of DARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

**Ethical Considerations**

Innovations in technology often face the ethical dilemma of dual use: the same advance may offer potential benefits and harms. For the IE technologies introduced in this tutorial, the distinction between beneficial use and harmful use depends mainly on the data. Proper use of the technology requires that input text corpora, as well as other modalities of inputs, are legally and ethically obtained. Regulation and standards provide a legal framework for ensuring that such data is properly used and that any individual whose data is used has the right to request its removal. In the absence of such regulation, society relies on those who apply technology to ensure that data is used in an ethical way. Besides, training and assessment data may be biased in ways that limit system accuracy on less well represented populations and in new domains, for example causing disparity of performance for different sub-populations based on ethnic, racial, gender, and other attributes. Furthermore, trained systems degrade when used on new data that is distant from their training data. Thus questions concerning generalizability and fairness need to be carefully considered when applying the IE technologies to specific datasets.

A general approach to ensure proper, rather than malicious, application of dual-use technology should: incorporate ethics considerations as the first-order principles in every step of the system design, maintain a high degree of transparency and interpretability of data, algorithms, models, and functionality throughout the system, make software available as open source for public verification and auditing, and explore countermeasures to protect vulnerable groups.
References


Brian Chen, Xudong Lin, Christopher Thomas, Manning Li, Shoya Yoshida, Lovish Chum, Heng Ji, and Shih-Fu Chang. 2021b. Joint multimedia event extraction from video and article. EMNLP Findings.


Joshua Feldman, Joe Davison, and Alexander M. Rush. 2019. Commonsense knowledge mining from pre-trained models. EMNLP.


Lifu Huang, Kyunghyun Cho, Boliang Zhang, Heng Ji, and Jiawei Han. 2021a. Cross-media structured common space for multimedia event extraction. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP2021).


Guoliang Ji, Kang Liu, Shizhu He, and Jun Zhao. 2017. Distant supervision for relation extraction with sentence-level attention and entity descriptions. In AAAI.


Manling Li, Sha Li, Zhenhailong Wang, Lifu Huang, Kyunghyun Cho, Heng Ji, and Jiawei Han. 2021a. The future is not one-dimensional: Complex event schema induction via graph modeling for event prediction. In Proc. The 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP2021).


Sha Li, Liyuan Liu, Yiqing Xie, Heng Ji, and Jiawei Han. 2022c. Piled: An identify-and-localize framework for few-shot event detection. In arXiv:2202.07615.


Xiaoya Li, Fan Yin, Zijun Sun, Xiayu Li, Arianna Yuan, Duo Chai, Mingxin Zhou, and Jiwei Li. 2019b. Entity-relation extraction as multi-turn question answering. In ACL.


Yaojie Lu, Hongyu Lin, Jin Xu, Xianpei Han, Jialong Tang, Annan Li, Le Sun, Meng Liao, and Shaoyi Chen. 2021. Text2event: Controllable sequence-to-structure generation for end-to-end event extraction. In ACL.


Qiang Ning, Hao Wu, Haoruo Peng, and Dan Roth. 2018b. Improving temporal relation extraction with a globally acquired statistical resource. In NAACL.


Zihan Wang, Jingbo Shang, Liyuan Liu, Lihao Lu, Ji-acheng Liu, and Jiawei Han. 2019b. CrossWeigh: Training named entity tagger from imperfect annotations. In *EMNLP*, pages 5154–5163, Hong Kong, China.


Pengfei Yu, Zixuan Zhang, Clare Voss, Jonathan May, and Heng Ji. 2022. Event extractor with only a few examples. In Proc. NAACL2022 workshop on Deep Learning for Low Resource NLP.


Xiyang Zhang, Muhao Chen, and Jonathan May. 2021b. Salience-aware event chain modeling for narrative understanding. In EMNLP.


Yunyi Zhang, Jianming Shen, Jingbo Shang, and Jiawei Han. 2020c. Empower entity set expansion via language model probing. ACL.


Ben Zhou, Daniel Khashabi, Chen-Tse Tsai, and Dan Roth. 2018. Zero-shot open entity typing as type-compatible grounding. In EMNLP.

Ben Zhou, Qiang Ning, Daniel Khashabi, and Dan Roth. 2020. Temporal common sense acquisition with minimal supervision. ACL.

Ben Zhou, Kyle Richardson, Qiang Ning, Tushar Khot, Ashish Sabharwal, and Dan Roth. 2021a. Temporal reasoning on implicit events from distant supervision. NAACL.


Wenxuan Zhou and Muhao Chen. 2021. Learning from noisy labels for entity-centric information extraction. In EMNLP.


A Appendix

A.1 Past Tutorials by the Instructors

The presenters of this tutorial have given the following tutorials at leading international conferences and venues in the past:

- Muhao Chen:
A.2 Recommended Paper List

The following is a reading list that could help provide background knowledge to the audience before attending this tutorial:

• Hangfeng He, Mingyuan Zhang, Qiang Ning, Dan Roth. Foreseeing the Benefits of Incidental Supervision. EMNLP, 2021.

• Ben Zhou, Qiang Ning, Daniel Khashabi, Dan Roth. Temporal Common Sense Acquisition with Minimal Supervision. ACL, 2020.


• Bangzheng Li, Wenpeng Yin, Muhao Chen. Ultrafine Entity Typing with Indirect Supervision from Natural Language Inference. TACL, 2022.


• Manling Li, Alireza Zareian, Qi Zeng, Spencer Whitehead, Di Lu, Heng Ji, Shih-Fu Chang. Cross-media structured common space for multimedia event extraction. ACL, 2020.