Robust, Generalizable and Minimally Supervised Knowledge Acquisition from Natural Language Text

Muhao Chen
Assistant Research Professor of Computer Science, USC

1 Introduction and Background

My research seeks to advance intelligent systems by developing the tools and techniques required for capturing world knowledge from natural language. Knowledge that expresses relations of concepts, entities and events, or interactions between objects in nature (such as molecules and biomolecules), is the backbone of any knowledge-driven AI systems and tasks. For example, narrative and dialogue processing systems incorporate procedural knowledge about events in order to anticipate what may happen next; e-commerce systems capture relational knowledge about products, users and user behaviors so as to effectively make recommendations; Web search systems query background knowledge representations to realize intelligent ways of content understanding and delivery; biomedicine research uses structural knowledge about proteins, drugs, diseases and clinical events to tackle high-stakes tasks such as drug-target detection, drug repurposing and high-risk disease early-diagnosis. Behind all those tasks, knowledge acquisition is the essential process that allows the machine to recognize actionable knowledge from the language we use, and further achieve synergetic knowledge sharing between systems, people and environments.

Despite the importance, research on knowledge acquisition has been facing critical challenges in the past decades. These challenges mainly lie in three aspects. First, obtaining supervision data for knowledge extraction is always a very expensive process that demands well-trained human experts to identify complex structures through reading long pieces of text. Accordingly, a reliable knowledge acquisition model should be achievable without the reliance on rich and high-quality supervision. Second, knowledge, even captured independently from separate sources, is often not isolated. For example, knowledge expressed in different languages may be complementary and unequally distributed [65]; knowledge stored in Gene Ontologies provides evidence for inferring interactions between new diseases and human proteins [20]. Therefore, knowledge acquisition has to support consolidation and transfer of knowledge across domains and resources. Third, the process of knowledge acquisition needs to generalize well, allowing portable deployment of knowledge in a wide scope of cross-domain, cross-modality and interdisciplinary tasks.

My research focuses on addressing these key challenges, aiming at developing robust, generalizable and minimally supervised machine learning methods for knowledge acquisition from natural language text. My investigations in this research area produced over 60 publications in leading AI, NLP and machine learning venues, documenting novel techniques that were deployed in many real-world applications, and delivered significant benefits in various computational and interdisciplinary areas. In the rest of this statement, I will briefly summarize my recent investigations and directions, illustrated in the research roadmap shown in Fig. 1. Then I will outline some of the exciting directions I plan to pursue in my future research.

Robust and Minimally Supervised Knowledge Extraction Structural knowledge representations, such as knowledge graphs (KGs), represent an integral part of intelligent systems in nearly all domains. For example, conceptual KGs like ConceptNet, ATOMIC and ASER have supported various narrative understanding and commonsense reasoning tasks; product and POI KGs provide evidential knowledge about consumer intents that have been essential to intelligent e-commerce technologies; biological and medical KGs, such as STRING, the Gene Ontology and DrugBank, have been supporting in silico research for vaccine development, drug repurposing and drug side effect detection. Despite their importance, existing efforts for acquiring structural knowledge representation are generally conducted manually. This leads to the cost of tens of millions of dollars to produce just one mid-scale commodity KG in the general domain, and this cost can be significantly higher in scientific domains like biology and medicine. To alleviate these costly efforts and provide reliable systems for automatically acquiring knowledge, my research has focused on robust and minimally supervised learning.
and inference technologies for knowledge extraction from natural language text. Specifically, my research systematically leads to transformative advancements in the following four dimensions.

(1) **Indirect supervision.** Learning to extract structural knowledge has largely relied on direct supervision from structurally annotated corpora that are similarly expensive as a structural knowledge representation itself [32]. I instead investigate a novel direction of indirect supervision, leading to robust and generalizable knowledge extraction models without sole reliance on task-specific annotations. In particular, my study has produced principled approaches for reformulating and transferring supervision signals from natural language inference (NLI) [25], summarization [29] and linguistic pattern matching [33]. This reformulation allows rich (indirect) supervision signals to be transferred from well-developed learning resources and models for signal-providing tasks that align well with knowledge extraction. It also emancipated knowledge extraction from the limitation of fixed label sets, allowing the inference of new types of knowledge that were unseen in training. In this context, I have also explored with semantic representation of tasks and labels [21, 14] to further reduce the need of direct task supervision. This systematic study of indirectly supervised learning has led to SOTA performance on a large number of benchmarks for relation extraction, named entity recognition, ultra-fine entity typing, event extraction and event process typing. Specifically, for all those tasks, my systems have demonstrated excellence in extremely few-shot [29, 21, 5] or zero-shot performances [25] that were close to those that perviously offered by full-shot, directly supervised models.

(2) **Noise- and perturbation-robustness.** In addition to insufficiency of annotated data, the cost and difficulty of structural annotation often lead to significant training noise. In the same context, real-world application scenarios often expose the model with way larger and more diverse data, for which the inference of model needs to frequently handle perturbations and out-of-distribution (OOD) exceptions. My study accordingly enhance the robustness of the model from two perspective. Towards robust training, my study developed a co-regularized knowledge distillation approach that can proactively identifying noisy training instances and preventing the discriminative model from fitting the noise [63]. This leads to significant improvement in both noise-robustness and computational efficiency over previous ensemble-based denoising and noise-filtering methods. In this context, my study also proposed sharpness-aware minimization with dynamic reweighting ($\delta$-SAM [66]) to further
enhance the model robustness using adversarial perturbation training, as well as self-supervised cross-lingual perturbation training [41]. On the other hand, to enhance the robustness in inference, I have studied margin-based contrastive learning methods [64, 7] that led to near-perfect unsupervised OOD detection performance, helping the model selectively identify cases where no extraction should be made. I also developed structure-aware equivariance learning techniques [44, 49] to allow data-to-text generation models to generate consistent representation for structural priors where semantic-invariant perturbations are free to be introduced. Those technologies systematically improves the reliability of knowledge extraction systems in real-world scenarios where training and inference phases are abundant with noise, perturbations and exceptions.

(3) Logically constrained learning and inference. Extracts are not standalone and can possess complex logical dependencies. A robust knowledge extraction system needs to ensure that the extracts are self-contained, and free of inconsistency and redundancy. My work accordingly suggests solutions to this problem with novel constrained learning and inference approaches. Specifically, I have studied joint constrained learning approaches for enforcing logical consistency in relation extraction tasks [45], probabilistic constrained learning with $t$-norm based optimization [18], logically constrained learning for linear relational embeddings [11] and probabilistic box embeddings [16]. Considering that logical constraints may be costly to define and hard to articulate, my recent study also proposed the approach to learn linear inequalities for automatically capturing logical constraints from data [46].

(4) Faithfulness. Current knowledge extraction models are mainly developed on large pre-trained language models and are short of training annotations in general. In this situation, my study has discovered that pre-training knowledge, distribution biases or existing annotation artifacts could often cause models to unfaithfully extract what is described in a given context, but instead to “guess” with a context-irrelevant extract using prediction shortcuts [50, 52]. Faithfulness, while being an under-explored research area, is undoubtedly a premise of reliable information extraction. In this context, my study has so far delivered several pilot studies to mitigate prediction shortcuts in entity-centric and event-centric information extraction with counterfactual analysis [50, 47], and counterfactual data augmentation [52]. On the other hand, to ensure that models make selective decisions on exception cases where nothing should be extracted, I contribute with selective prediction techniques based on high-order metric learning [28, 35] and Dirichlet parameterization [47].

My main contributions in this line of research were summarized in the tutorials I presented at NAACL 2022 [3], ACL 2021 [12] and other invited talks, and led to the support I am receiving from the DARPA KMASS program, the DARPA MCS program, and the Cisco Faculty Research Award.

Transferable Representation Learning for Structural Knowledge. Structural representation learning is the requisite for incorporating symbolic knowledge into deep learning models. A key contribution I have made to this field is on the transferability of such representations. Different domains or sources of data, or even different languages, often provide interchangeable and complementary knowledge. Hence, it is particularly important to develop a universal representation learning method that captures the association of knowledge across multiple data sources with minimal supervision, and support with credible knowledge transfer across different domains. I started this line of research and provided the first embedding framework that bridges multiple language-specific KGs [10, 15], by performing semi-supervised alignment of multiple relational embedding models. To more precisely capture the knowledge association with minimal supervision, I have extensively extended the alignment learning process based on iterative co-training [8], multi-view representation [9, 56], incidental supervision from free text [6], unsupervised visual pivoting [27] and coarse-to-fine entity linking [19]. I have also devised relational embedding techniques that are robust against scarcity and structural heterogeneity of data, using techniques based on box embeddings [16], concept contextualization [37] and attentive neighborhood aggregation [38]. Particularly, for highly complex knowledge-representation structures, I devised on new paradigms for non-linear embedding spaces [36, 30, 5]. For knowledge transfer from multiple sources of (inconsistent) learning resources, my work addresses the problem of inducing trustworthy inference results with ensemble knowledge transfer [17, 65]. In this context, my study also contributes with answer consolidation [67] and multi-modal fact verification techniques [43] that help resolve the redundancy and inconsistency.

muhaoche@usc.edu

https://luka-group.github.io/
of local extracts for global knowledge representation.

This line of research has received a wide recognition by the community, and the importance of this contribution has been recognized by over a thousand citations collectively in the past four years. A wide spectrum of applications have also been benefited from the techniques proposed in my papers and follow-up works. The advancement in this research topic has been featured in my tutorial at AAAI-2020 [2] and our recent benchmarking paper [39], and has been recognized with an NSF CRII Award in 2021.

**Portable Knowledge-driven Applications** The robust knowledge extraction and knowledge transfer technologies allow efficient and broad utility of the knowledge in computational research. This allows my study to further enhance the various narrative understanding [59, 42] and commonsense reasoning tasks [54, 14, 34, 33, 49, 48] that are at the core of current NLP research. My recent study also extends the utility of robust information extraction [60] and summarization techniques [23, 24] for knowledge acquisition and dissemination from online programming tutorials and discussions, aiming at helping software developers in making informed decisions and supporting the development of knowledge-based tools. Besides, I also gained critical experience on how to best transfer the above two lines of technologies to tackle important tasks in computational biology and medicine, including protein-protein interaction prediction [4], proteomic mutation effect estimation [61], circular RNA detection [22], disease target identification [20], and clinical diagnostic prediction [57, 58]. In addition, a recent thread of my research also interfaces language modeling with geographical data [26], aiming at benefiting automated machine comprehension and dissemination of digital map information. I am excited about the broad utility of knowledge acquired from my fundamental research on information extraction and knowledge transfer, and will continue the investigation on solving important problems in various related areas.

2 **Research Agenda: Directions for Future Work**

My future research will continue to focus on improving machines’ ability of efficiently and reliably acquiring general-domain and expert knowledge with minimal supervision, and leverage transferable representations to solve problems in various domains and interdisciplinary areas. Particularly, I intend to extend my investigation in the following directions.

**Event-Centric NLP.** Human languages evolve to communicate about events happening in the real world. Therefore, understanding events plays a critical role in natural language understanding. A key challenge in accomplishing this mission lies in the fact that events are not just simple, standalone predicates. Rather, they are often described at different granularities, temporally form procedures, and are always directed by specific goals and sub-goals. Understanding events requires the understanding of how events are connected, form procedural or membership structures, and recognizing typical properties of events (e.g., space, time, salience, essentiality, implicitness, preconditions, consequences, etc.). Our prior studies have focused on inferring the logical constraints [45, 46], analogical properties [54], evolution patterns [68], salience [59] and membership relations [14] of events. Yet, more fundamental challenges persist. Particularly, unlike entities, events have an internal structure, whereby ensuring consistent information extraction for internal components and external relations of events remain an unresolved problem. Besides, machine comprehension of causal relations, conditions and consequences requires unrealized human-like cognitive understanding. As events may evolve and be described in different ways in different documents, inferring the relations of events across documents, consolidating highly diverse cross-document event descriptions into unique and consistent global knowledge representation represents a daunting challenge to event-centric information extraction. Moreover, reasoning about events challenges commonsense reasoning agents with common implicitness of event mentions, event arguments and event properties (e.g., essentiality [55], preconditions [33], and spatial attributes). My recent tutorials at ACL 2021 [12] and AAAI 2021 [13] have systematically summarized the current (pre-mature) status of event-centric NLP, and positioned the emerging fundamental research challenges, including the above, that will be at the core of my future research. In this context, I also plan to investigate the use of event knowledge to improve the coherence of narrative (e.g., our prior study [59]) and dialogue systems, enhance the factuality of summarization systems, as well as realize clinical event understanding to tackle important but expensive
clinical diagnostic tasks [58, 57].

Harnessing Massive Language Models. Massive language models (also known as foundation models) such as GPT-3 and PaLM have excelled in many NLP tasks with their strong ability of distilling knowledge from Web-scale pre-training corpora, and raised opportunities in research directions such as in-context learning. Despite the progress, there are two critical issues with NLP solutions based on massive language models. First, massive language models still fall short of supporting reasoning, including logical, quantitative and cognitive reasoning. Our recent study on machine comprehension of medical reports and tabular data has proven that massive language models still fail to correctly infer the relations between time intervals, and do not meaningfully support numeracy. Our study also found that GPT-3 fails at many cognitive understanding tasks. For example, they achieve an AUC of 63% when inferring the essentiality of events [55], whereas humans can achieve around 87%. Second, after costing tens of millions of dollar to be trained, massive language models have to remain fixed within their year-long life cycles, causing them to be inadaptive to vastly streaming new information about the ever-changing world. My future research will tackle these issues in two directions. First, I will develop various mid-scale language models where dedicated kernel functions or neural symbolic modules are incorporated into the Transformer architectures, seeking to realize diverse types of reasoning processes. These models will form mixture-of-experts (MoE) or hero-gang structures together with the massive language model, providing complementary reasoning abilities. Second, towards adaptation to changing information, I will mainly investigate two methodologies. Specifically, I will continue our study on parameter-efficient adaptation [53] to allow plug-in of memory modules about new knowledge into the MoE structure, for which added information in the memory modules may come from both new data and human-in-the-loop. Moreover, I will leverage our robust information extraction technologies to timely capture new information from the Web, and realize retrieval-augmentation of massive language models at inference.

Generalizing Indirect Supervision for Scalable NLP. My prior studies have demonstrated the success of indirect supervision from NLI and NLG tasks in enhancing information extraction within the scope of single documents [25, 21, 29]. I will further extend this line of study in three directions. First, to allow more enriched knowledge extraction, it is essential to enable the extraction process across documents. As cross-document extraction tasks, such as cross-document co-reference resolution and cross-document relation extraction suffer more severely from insufficient training data than their single-document counterparts, my study will focus on (i) multi-hop dense retrieval approaches [51] for discovering evidence that supports cross-document relations, and (ii) indirect supervision from multi-document tasks, such as multi-document summarization and sentence fusion. Second, beyond the use of existing source tasks for supervision signals, I will also design linguistic pattern mining and generative data augmentation methods that can automatically find or generate large-scale weak supervision data (a preliminary study has been conducted for preconditioned commonsense inference [33]). In the same context, I will further study methods that proactively select and filter task-specific weak supervision data following my prior studies on unsupervised denoising [63, 66]. Last, to understand the learnability of different sources of indirect supervision, I will study affinity measurements to quantify the contribution made by each source task to the target task objectives from both data and model parameter perspectives, and ways to capture and represent logical constraints that should be enforced among cross-task decisions.

Equivariance Learning in NLP. As an important but largely under-explored component of robust NLP systems, both the language understanding and generation processes need to handle equivariance properties in data. For example, the narrative structure of an article can be reorganized, while still presenting the same content. In constrained NLG tasks with structural priors (e.g. structured data-to-text generation), the structure of the prior can also be modified while presenting semantically equivalent content. However, existing sequential modeling of languages cause downstream information extraction and NLG systems to be brittle to content-neutral transformations of input data. Our pilot study realizes equivariance learning by incorporating structured masking and transformation-invariant position encoding mechanisms in pre-trained Transformer models for data-to-text [44] and scene-to-text [49] generation tasks. Following this direction, I will investigate principled approaches
to address three research questions: (i) How to capture implicit but semantically equivalent structures (e.g., narrative structures) of natural language text, and accordingly generate equivariant language representations; (ii) How to disentangle semantic and syntactic representation in large language models; (iii) How to composite information from multiple components of text (e.g., sentences, paragraphs, or documents) while ensuring the equivariance to positional and frequentional perturbations.

Cross-domain and Interdisciplinary Research. I always believe that a useful technology should address problems in several related research areas rather than a single one. Therefore, beyond core NLP tasks, my research has also contributed to computational biology [4, 61, 22], medical informatics [58, 57, 20], software engineering [23, 24, 60], geo-intelligence [26] and social media analysis [40, 1]. Particularly, I have been interested in AI technologies for common good that could contribute to fairness [62], healthcare [58, 57] and education [31]. Given the previous success in transferring technologies to different areas, I am enthusiastic about developing open-source libraries and software, and facilitating collaborations with people outside my areas. I am excited about any opportunities to apply my expertise in NLP and representation learning to solve important problems in other areas and disciplines.

References


muhaocche@usc.edu 6  https://luka-group.github.io/


[34] QASEMI, E., LIU, F., CHEN, M., AND SZEKELY, P. Paco: Preconditions attributed to commonsense knowledge. In EMNLP (2022), in submission.


[38] SUN, Z., WANG, C., HU, W., CHEN, M., DAI, J., ZHANG, W., AND QU, Y. Knowledge graph alignment network with gated multi-hop neighborhood aggregation. In AAAI (2020).


[63] Zhou, W., and Chen, M. Learning from noisy labels for entity-centric information extraction. In
EMNLP (2021).


