Knowledge Acquisition with Transferable and Scalable Representation Learning

1 Introduction and Background

My research focuses on promoting the advancement of intelligent computational systems with better awareness of commonsense and expert knowledge about the world, which lead to more efficient information exchange of the system with human and the nature. My goal, in the long run, is to leverage unified methodologies to help machines understand the relations of lexemes, entities and concepts in human languages as well as apprehend the interactions of objects (such as molecules and biomolecules) in nature. In the near term, I am motivated by the development of new technologies in representation learning and information extraction, and extending their benefits in various tasks for knowledge base construction, natural language understanding and computational biology.

The challenges span in a range of fields, from the fundamental questions in the representation, learning and inference of knowledge, to systematic paradigms for scalable data management, mining and retrieval. In this decade, AI systems in various application domains are empowered by representation learning technologies for automatically discovering and acquiring relations, patterns and properties of objects from large-scale data. In particular, such technologies involve multi-relational embeddings, language modeling and sequence pair modeling. My work seeks to trigger the advancement of these technologies, focus on knowledge acquisition from data in different modalities, and in scenarios with and without plausible supervision signals.

During my Ph.D. studies, I have focused on exploring over a number of interrelated research questions:

1. Can systems discover and capture the inherent associations of complementary and interchangeable knowledge described by different data, therefore helping the integration and synchronization of knowledge across different domains and data sources?
2. Can representation learning models support the learning and inference for structured data with complex properties (for example, uncertainty, logic properties and hierarchies)?
3. Can deep learning techniques for languages effectively and efficiently capture relations from large-scale corpora, while being robust against certain bias in domain-specific data? Moreover, can these techniques be extended beyond human languages, such as genomic data, proteomic data, programming languages and electronic health records?

My investigation on these key research questions have led to over 30 published papers, and have benefited real-world applications in various computational and interdisciplinary areas. The following of this research statement presents parts of my investigation in the past four years into these questions, followed some exciting future directions I am willing to expand my research towards.

Transferable Representation Learning for Multi-relational Data Multi-relational data, such as knowledge graphs (KGs), lexical graphs, product graphs and biological networks, serve as the essential source of both commonsense and domain knowledge for various knowledge-driven AI systems. Such data involve the curation of semantic relations of various objects, entities, concepts and lexemes as well as diverse types of interactions between molecules and biomolecules in the nature.

My research has focused on multi-relational representation learning, which seeks to support the characterization and inference of relational knowledge in embedding spaces. This is the requisite to incorporate the symbolic knowledge into deep learning models. Specifically, a key contribution I have made to this community is on the *transferability* of multi-relational representation learning. Different sources of multi-relational data, often provide *interchangeable* and *complementary* domain knowledge. Hence, it is particularly important to develop a general representation learning method that is able to capture the association of knowledge across multiple data sources, and further support the transfer and synchronization of knowledge across different domains. I started this line of research by proposing the first multi-relational embedding framework

that bridges multiple language-specific KGs [12, 16], which is based on joint learning of multiple relational embedding models and a semi-supervised alignment learning model. To more precisely capture the knowledge association with minimal or almost no supervision, I have extensively extended the alignment learning process based on iterative co-training [10], multi-view learning [28] and incidental supervision [8]. I have also contributed with relational embedding techniques that are robust against scarcity and structural heterogeneity of data, particularly based on relation contextualization [25], attentive neighborhood aggregation [26] and hyperbolic representation learning [24]. For inference with transferable representations of inconsistent datasets, I also address the problem of inducing trustworthiness of inconsistent prediction based on ensemble inference [17].

This line of research has seen wide recognition by the community for its importance, and has received over a hundred citations in the past two years. It has also seen a wide spectrum of applications benefited from proposed techniques. Related techniques also reach state-of-the-art on few-shot entity typing [19] with knowledge transfer between commonsense ontologies and instance KGs, and effectively improves protein-protein interaction prediction and single-cell RNA expression imputation with knowledge transfer from gene ontologies [2]. The advancement in this research topic will be systematically summarized in a part of our upcoming tutorial at AAAI-2020 [3].

Relation Induction from Large-scale Complex Data While multi-relational data have crucially provided structural and actionable knowledge representations to many AI applications. the construction of such data is costly and has often relied on extensive human efforts. Hence, it is an urged demand to develop methods that automatically acquire relational knowledge from unstructured data.

My research has developed data-driven relation induction methods from massive data of other modalities. Under this topic, I have studied *large-scale document relation extraction*. My collaboration with Google Knowledge Graph has delivered a scalable system for document relation induction [6]. The system has received a near-perfect of over 98% F1 score for *subtopic* relation extraction on a ten-thousand-scale annotated data. Based on distributed inference with a thousand-machine MapReduce, the system was able to efficiently extract relations from 24 trillion Wikipedia article pairs within several hours. This research project also delivered by-products for entity extraction [4] and article readability analysis [21]. My study has also delivered competitive end-to-end systems for *interaction prediction of biomolecules*, which seeks to populate biological data banks and provide effective *in silico* techniques to alleviate the laborious and expensive *in vivo* or *in vitor* efforts from biologists. These include the first sequence-based end-to-end system for multifaceted Protein-protein interaction (PPI) prediction [5], and the point mutation effect estimation method based on contextualized representation learning for nucleotides and multi-task learning [29]. The proposed methods in this thread of research has so far consistently offered state-of-the-art performance on various tasks such as PPI link prediction, PPI type prediction, change of binding affinity estimation and buried surface area estimation. In this line of research, I have also studied paraphrase-aware contextualized word embeddings for robust discourse relation detection [9], and jointl learning of lexical and sentential semantics for semantic search [11, 22].

Learning and Inference with Complex Properties Many data have highly complex properties and structures. In such situations, representation learning and inference of knowledge is often intractable. My work accordingly suggests solutions to such problems from several perspectives. In [18], I have proposed a novel uncertainty-aware representation learning method to capture both relational structures and uncertainty with embeddings, which incorporates the proactive Probabistic Soft Logic reasoning process with the embedding learning process, and aims at precise inference of the confidence for unseen relation facts. This approach has been successfully applied to drug discovery [23]. I have designed an approach to preserve hierarchical relational structures with complex profile information of relations in hyperbolic spaces [7]. I have also extended the aforementioned approach to achieve state-of-the-art performance on knowledge alignment and entity typing tasks [24]. In [19], my work has also demonstrated the potential of these representation learning

approaches to improve hierarchical classification and extreme classification problems with better inference ability of the label space. In this line of research, I also studied approaches for capturing logical properties of relations in linear embedding spaces [13] and supporting multi-label learning on multigraphs [1].

2 Research Agenda: Directions for Future Work

My research goal is to help machines acquire commonsense and expert knowledge through unified methodolgies of representation learning, and leverage transferable knowledge representations to solve problems in various domains and interdisciplinary areas. I have already made significant strides toward this goal, and intend to focus on further extension of the methodologies described above over the next few years. In the near future, I am excited to work on the following directions.

Robust Machine Learning with Generalizability. Although data-driven machine learning benefits with automatic feature extraction, end-to-end training and inference, the produced models often suffer from generalization issues. Hence, I am interested in developing robust learning systems for various domain tasks. The methods I seek to explore include massively pre-training representation learning models and supporting domain invariant feature extraction. Through the investigation, the proposed techniques will be applied to a broad range of cross-lingual and low-resource language tasks. Meanwhile, corresponding techniques are also urged by the computational biology community to enable tractable cross-species prediction tasks. A recent study [20] has successfully used our PIPR model [5] trained on Arabidopsis to infer protein functions on tomato, while finding the transferability from yeast to Arabidopsis to be quite limited (the later two have a $10\times$ evolution distance in comparison to the former two). Given the intractable cross-species transferability of current methods, my future research would seek to deliver multi-species M-BERT-like models and species-invariant adversarial learning models. This is with the purpose of generally benefitting the computational biology community that studies on genomic and proteomic data. On the other hand, I also seek to improve the inference of models by addressing the "null prediction" problem, which aim at teaching the model to understand the cases where there is no correct answers. I would like to tackle this challenge by developing a general adversarial learning framework to help a task-specific model to discriminate between answerable and unanswerable cases, and explore with advanced ordinal metric learning techniques to improve the inference ability of the model.

Learning with Minimal and Incidental Supervision. Training data-driven machine learning models necessarily require extensive collection of learning resources. However, annotating task-specific outputs is difficult and often requires expertise and domain knowledge, making it costly to obtain annotations that with high quality and fairness. On the other hand, the internet has produced enormous amounts of textual and multimodal data. Although these data may not directly come with annotations for specific end tasks, they involve related information that can be transformed into incidental supervision signals. However, the massive volume and noisy nature of unstructured and multimodal data present challenges to existing learning algorithms. Based on my pilot study of transferable representation learning with incidental supervision [8], I seek to further tackle the technical challenges of exploiting implicit and analogous supervision signals from heterogeneous learning resources by designing effective self-supervised learning algorithms. Towards this end, I am also interested in incorporating transferable representation learning of symbolic knowledge into learning systems, so as to support the systems with better inference ability on tasks with insufficient, biased training data or extreme label spaces.

Non-linear Characterization of Highly Complex Structures. Many applications involve highly complex structures, which are particularly difficult for linear representation learning models to capture, I am interested in investigating new paradigms of feature modeling with non-linearity. Based on my previous study on hyperbolic representation learning for multi-relational data [7, 24], I would like to enable learning systems with non-linear characterization of complex objects that form hierarchies, such as the evolution of claims and programs in online social media and software repositories. I am also interested in deploying efficient

set learning mechanisms to help systems seizing naturally unordered data, such as medical events in EHR data of each single patient visit.

Language Modeling with Awareness of Knowledge. Language models are the backbone of many NLP systems. They critically capture the statistical relations of language components, and provide contextualized semantic representations for lexemes and sentences. While language models are mostly trained from scratch on raw text corpora, a line of research that has always been in my mind is to incorporate language models with both linguistic and commonsense knowledge [14]. My previous study has achieved robust contextualized representations of lexemes and sentences by making language models aware of contextual paraphrasing [9]. Specifically, I believe a more reliable semantic representation method should also consider other linguistic properties of contexts, including different discourse relations and semantic classification of the linguistic contexts. On the other hand, various knowledge-driven NLP tasks would tremendously benefit from a language model that is aware of commonsense knowledge. Through this line of research, I intent to study how we can probe language models with multiple aspects of relational knowledge, lexicographic knowledge [11] and temporal knowledge.

Cross-domain and interdisciplinary research. I always believe that a useful technology should not be limited by problems in a single research area. As presented in this research statement, my research for multi-relational representation learning and relation induction has contributed to a broad range of tasks in knowledge base construction [12, 6], natural language processing [11, 9], computational biology [5, 29], computer networks [15] and social media analysis [27, 21]. Some of my ongoing or planned works are also extending the proposed methods in my Ph.D. study to new application scenarios, including event prediction on EHR data with knowledge transfer from disease ontologies, and projecting protein and disease knowledge to predict interactions of drugs. Given the previous success in transfering technologies to different areas, I am enthusiastic about developing open-source source libraries and software, and facilitate collaborations from people outside my areas. I am excited about any opportunities to apply my expertise in representation learning and NLP to solve important problems in other areas and disciplines.

References

[1] CHEN, H., TIAN, Y., CHEN, M., PEROZZI, B., AND SKIENA, S. Enhanced network embeddings via exploiting edge labels. In *CIKM* (2018).

[2] CHEN, M. Multi-relational representation learning and knowledge acqusition. *UCLA Doctoral Dissertation* (2019).

[3] CHEN, M., CHANG, K.-W., AND ROTH, D. Recent advances in transferable representation learning. In *AAAI Tutorials* (2020).

[4] CHEN, M., HUANG, G., AND ZANIOLO, C. Learning to differentiate between main-articles and sub-articles in wikipedia. In *BigData* (2019).

[5] CHEN, M., JU, C., ZHOU, G., CHEN, X., ZHANG, T., CHANG, K.-W., ZANIOLO, C., AND WANG, W. Multifaceted protein-protein interaction prediction based on siamese residual rcnn. *Bioinformatics 35*, 14 (07 2019), i305–i314 (Procs of ISMB 2019).

[6] CHEN, M., MENG, C. P., HUANG, G., AND ZANIOLO, C. Neural article pair modeling for wikipedia sub-article machine. In *ECML* (2018).

[7] CHEN, M., AND QUIRK, C. Embedding edge-attributed relational hierarchies. In *SIGIR* (2019).

[8] CHEN, M., SHI, W., ZHOU, B., AND ROTH, D. Cross-lingual entity alignment for knowledge graphs with incidental supervision from free text. In *ACL (In review)* (2020).

[9] CHEN, M., SHI, W., ZHOU, P., AND CHANG, K.-W. Retrofitting contextualized word embeddings with paraphrases. In *EMNLP* (2019).

[10] CHEN, M., TIAN, Y., CHANG, K.-W., SKIENA, S., AND ZANIOLO, C. Co-training embeddings of knowledge graphs and entity descriptions for cross-lingual entity alignment. In *IJCAI* (2018).

[11] CHEN, M., TIAN, Y., CHEN, H., CHANG, K.-W., SKIENA, S., AND ZANIOLO, C. Learning to represent bilingual dictionaries. In *CoNLL* (2019).

[12] CHEN, M., TIAN, Y., ET AL. Multilingual knowledge graph embeddings for cross-lingual knowledge alignment. In *IJCAI* (2017).

[13] CHEN, M., TIAN, Y., ET AL. On2vec: Embedding-based relation prediction for ontology population. In *SDM* (2018).

[14] CHEN, M., AND ZANIOLO, C. Learning multi-faceted knowledge graph embeddings for natural language processing. In *IJCAI* (2017).

[15] CHEN, M., ZHAO, Q., DU, P., ZANIOLO, C., AND GERLA, M. Demand-driven cache allocation based on context-aware collaborative filtering. In *MobiHoc* (2018).

[16] CHEN, M., ZHOU, T., ET AL. Multi-graph affinity embeddings for multilingual knowledge graphs. In *AKBC* (2017).

[17] CHEN, X., CHEN, M., FAN, C., UPPUNDA, A., AND ZANIOLO, C. Cross-lingual knowledge graph completion via ensemble knowledge projection. In *ACL (In review)* (2020).

[18] CHEN, X., CHEN, M., SHI, W., SUN, Y., AND ZANIOLO, C. Embedding uncertain knowledge graph. In *AAAI* (2019).

[19] HAO, J., CHEN, M., YU, W., SUN, Y., AND WANG, W. Universal representation learning of knowledge bases by jointly embedding ontological concepts and instances. In *KDD* (2019).

[20] MAKRODIMITRIS, S., VAN HAM, R., AND REINDERS, M. Sparsity of protein-protein interaction networks hinders function prediction in non-model species. *bioRxiv* (2019), 832253.

[21] MENG, C., CHEN, M., MAO, J., AND RIBEIRO, B. Readability analysis of web articles using hierarchical lstm. In *ECIR* (2020).

[22] SHI, W., CHEN, M., TIAN, Y., AND CHANG, K.-W. Learning bilingual word embeddings using lexical definitions. In *RepL4NLP* (2019).

[23] SOSA, D. N., DERRY, A., GUO, M., WEI, E., BRINTON, C., AND ALTMAN, R. B. A literature-based knowledge graph embedding method for identifying drug repurposing opportunities in rare diseases. In *PSB* (2020).

[24] SUN, Z., CHEN, M., HU, W., WANG, C., DAI, J., AND ZHANG, W. Knowledge association with hyperbolic representation learning of knowledge graphs. In *ACL (In review)* (2020).

[25] SUN, Z., HUANG, J., HU, W., CHEN, M., AND QU, Y. Transedge: Translating relation-contextualized embeddings for knowledge graphs. In *ISWC* (2019).

[26] 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).

[27] TIAN, Y., CHEN, H., CHEN, M., PEROZZI, B., AND SKIENA, S. Social relation inference via label propagation. In *ECIR* (2019).

[28] ZHANG, Q., SUN, Z., CHEN, M., GUO, L., AND QU, Y. Multi-view knowledge graph embedding for entity alignment. In *IJCAI* (2019).

[29] ZHOU, G., CHEN, M., JU, C., WANG, Z., JIANG, J.-Y., AND WANG, W. Mutation effect estimation on proteinprotein interactions using deep contextualized representation learning. *NAR Genom Bioinform* (2020).