Research Statement

Peng Qi
http://qipeng.me/

One of the most important applications for natural language processing (NLP) systems is to help fulfill humans' information needs using the knowledge in large collections of text. To achieve this goal, during my Ph.D., I have worked on research projects that help answer our questions from large text collections.

In recent years, the NLP community has made great progress on question answering (QA) systems. For instance, these systems can answer questions like *When was Albert Einstein born?* with factoid answers extracted from Wikipedia. However, they still largely rely on matching local patterns in text, and fall short at fulfilling information needs that require multiple steps of reasoning. For example, in a conversation about Albert Einstein, current systems struggle with questions like *Where was his father born?*, because this question requires an additional step of reasoning about who Einstein's father is. Moreover, the factoid answer (*Bad Buchau*) would not be very informative to most interlocutors, and does not allow the user to pursue information to a greater depth on this topic (in contrast, a human conversationalist would probably have answered *Bad Buchau, a small town in South Germany not far from Stuttgart and Munich*, which prompts more follow-ups). Enabling NLP systems to perform multi-hop reasoning and to provide appropriately informative answers in conversational QA have thus been two main directions of my research.

Multi-hop reasoning is crucial in allowing us to gain more insight into textual knowledge. I have attacked this problem from two distinctive angles. On the one hand, I built an accurate model that combines the latest developments in graph neural networks with linguistic insight to construct symbolic knowledge bases from text to enable multi-hop inference [3]. On the other hand, I have also worked on purely text-based multi-hop QA free from predefined knowledge schemas, by building one of the first multi-hop QA datasets [2] and developing an efficient and explainable QA system that outperforms previous work that utilizes much more powerful pretrained neural network components [1].

Complex information needs are also common in conversations, where factoid answers are usually insufficient. An ideal system should contribute additional information to the conversation when necessary to help the user acquire information in greater depth. To gauge how successful different answering strategies are, we need to first be able to evaluate how successful different answers are at prompting appropriate and meaningful follow-up questions in an information-seeking dialogue. As a first step, we built a model of an inquisitive interlocutor by generating questions in an information-seeking conversation [15]. This will not only afford future NLP systems a means to learn from humans, but more importantly, it will provide insight into their users' state of mind, and help these systems decide the right amount of information to offer in conversations.

Underlying these main directions, one common theme in my research has been model *explainability*, *i.e.*, models should be designed to explicitly demonstrate their decision process when possible, and to allow for intervention if necessary. It supersedes *interpretability*, which, although it improves transparency, leaves the burden of interpretation to us (*e.g.*, [5]), and comes with fewer guarantees of system behavior. I believe explanability is not only desirable for NLP systems to borrow human insight such as linguistic and quasi-linguistic observations as inductive bias, but also essential if we want to deploy them in the real world.

Multi-hop Reasoning: Building Knowledge Bases with Linguistic Insight

One efficient way to answer questions from large amounts of textual knowledge is to turn text into a knowledge base (KB), a collection of tuples that declare relations between pairs of entities (*e.g.*, [9, 8]). For instance, we can convert the sentence *Barack Obama was born on August 4, 1961 in Honolulu, Hawaii, not Kenya* into tuples such as *(Barack_Obama, born_in, Hawaii)* and *(Barack_Obama, date_of_birth, 8/4/1961)*. NLP systems can then perform explainable multi-hop reasoning through composing and aggregating over these tuples to answer complex questions like *Which U.S. President born after 1900 was not born in one of the lower 48 states?*

Traditionally, KBs are largely constructed either manually or with expert-written patterns such as regular expressions based on linguistic information such as parts-of-speech and syntactic trees. Although accurate, these approaches are slow and costly to scale up, and thus suffer from low coverage of the knowledge in text. While data-driven approaches can help mitigate this issue, previous work either uses shallow statistical learning methods or superficial features, and is thus often less robust than ideal.

We instead explored a combination of powerful deep graph neural networks with linguistic insight for more accurate automatic KB construction from text [3]. To better model the relation between entities, our model makes use of *dependency parses*, a kind of syntactic analysis of sentences I have worked extensively on [6, 7, 4, 16]. Because tree structures are not amenable to efficient batched linear algebra operations, previous work often reduced them to the shortest dependency path (SDP) between the entities. Although effective at filtering out irrelevant information in most cases, this simplification neglects crucial information sometimes. For instance, the SDP between *Barack Obama* and *Kenya* in our example sentence leaves out the negation (*"not"*), which results in catastrophic misclassification.

We make the linguistic observation that words closer to the SDP are more pertinent to discerning the relation between entities, and thus prune words that are farther away from the SDP. We combine this technique with graph convolutional networks (GCN), a class of neural networks tailored for efficient modeling of graphical structures, to learn dependency patterns that express entity relations from data. The resulting system was the state of the art on TACRED, one of the largest datasets for this task.

Multi-hop Reasoning: Open-domain Text-based Question Answering

Knowledge bases provide strong typing for entities and relations that can be easily composed for scalable inference, but this sometimes comes at the price of lacking flexibility. Adding new entities and relations to broaden the coverage of real-world knowledge bases is a non-trivial task. In contrast, text-based question answering (QA) systems rely more on semantic pattern matching, and can generalize to new textual knowledge more easily without well-defined schemas (*e.g.*, Wikipedia). However, previous text-based QA systems pale in comparison to KBs on multi-hop reasoning, which is a result of the QA datasets available to the research community.

To enable QA systems to perform complex reasoning, my collaborators and I collected one of the first multi-document multi-hop QA datasets, named HOTPOTQA [2]. We devised a new method for data collection to ensure the questions, collected on Wikipedia articles, are free from pre-defined KB schemas (*i.e.*, pre-defined relations between usually a dozen types of entities) and diverse in content and reasoning strategy. Borrowing the idea from symbolic multi-hop reasoning in KBs and ontologies, we make use of hyperlinks in Wikipedia for entity overlap, along with curated lists of common entities with shared properties (*e.g.*, Astronauts in Space) to produce pairs of related documents for efficient crowd-sourcing. We also designed the dataset for explainable QA systems, and ask crowd workers to annotate the sentences required to answer the questions as rationales for explanation. Systems can then be trained to predict these sentences as supporting facts for their answer. HOTPOTQA has fueled research in multi-hop QA in both a *few-document* setting where systems only process ten documents for each question, and a fully *open-domain* setting where they find the answer and supporting facts from the entire Wikipedia.

Open-domain QA is a very practical but challenging task in NLP, and performing multi-hop reasoning in a large collection of text even more so. The main difficulty lies in finding all the necessary pieces of evidence in this collection. Most previous open-domain QA systems focus on single-hop questions, where it suffices to search the text collection with the question itself. This strategy fails, however, for questions like *When was his father born?* when conversing about Albert Einstein, because the question does not provide sufficient context to retrieve supporting facts effectively.

We propose a solution to open-domain multi-hop reasoning with an iterative retrieve-and-read approach [1]. Our model, GOLDEN (Gold Entity) Retriever, iterates between reading the retrieved documents and generating natural language queries to retrieve more supporting facts to answer the question. In our example, the model first searches for *Albert Einstein*, then reads the top retrieved articles to search for *Hermann Einstein*, his father. Training these query generator models in an end-to-end optimized pipeline is intractable because the search space is prohibitively large. We solve this problem by making a minimal but universal observation: at each step of reasoning, there exists a strong semantic overlap between *the knowns* (question + retrieved documents) and *the wants* (evidence documents yet to be retrieved), which is an ideal query candidate. This avoids the need for expensive reinforcement learning during training, where the system has to explore different search queries end-to-end, and receive very sparse rewards for successfully answered questions. Instead, we turn the problem into one of *imitation learning*, where we have effectively derived an oracle that demonstrates a path towards the final answer through the vast search space. We further propose to generate search queries by extracting a contiguous span from the retrieved documents with a QA model, which yields coherent and explainable queries. The resulting system outperformed all previously published systems that used large, massively pretrained NLP models (*e.g.*, BERT) on this dataset, and remains one of the top systems while being more computationally efficient and explainable.

Effective QA in Conversations: Inferring Information-seeking Intent

In natural human-human interactions, questions never come up in isolation, nor are their answers devoid of context. For example, the question *How long is it going to take?* elicits very different answers in a conversation about restoration of the Notre Dame Cathedral and one about academic paper reviews. Furthermore, answering these questions as a good conversationalist also requires understanding the background of one's interlocutor. When talking about Notre Dame, offering detailed accounts of the fire damage aside from answering restoration length might be informative to someone how hasn't been following the news closely and helps open up the conversation, but the same would sound tedious and redundant if the interlocutor is well up-to-date on this matter.

To build effective and practical conversational NLP systems, it is important that we move away from thinking of conversation as a form of text, but rather as a means of information exchange. Thus, in our recent work, we take the first step towards catering answers to the information needs of the interlocutor in an information-seeking conversation. Instead of building more question answering systems or datasets, we approach this problem by attempting to understand the interlocutor's intent in such conversations in order to decide the right amount of information to provide in the conversation. We build a conversational question *asker*, which predicts what questions can be asked given a conversation history to model the behavior of an information seeker [15]. Rather than assuming access to potential answers as is standard practice in previous work on question generation, we instead turn to inducing and leveraging quasi-linguistic structures such as conversation structure to build an explainable and controllable system. Specifically, our model induces for each question in a QA conversation which question/answer pair(s) it depends on, which we show empirically improves the quality of questions generated, helps humans better understand the underlying structure of the conversation, and allows us to explore alternative questions by manipulating this dependency at evaluation time. This also allows future NLP systems to extrapolate interlocutor state of mind and intent in such information-seeking conversations, and tailor the answers provided to sustain more engaging conversations.

Future Directions

The work I have mentioned touches upon one of the core values of NLP systems—serving our information needs. Many open questions remain in this direction, but one of the most important problems I would like to focus my future research on is building explainability into artificial intelligence (AI) systems. Explainability is crucial in building trust and acting as a failsafe if AI systems are to be deployed in the real world, and this holds beyond the information-serving NLP systems I have worked on. Moreover, I believe that explainability is one of the most essential inductive biases that contribute to better data efficiency and robustness of statistical machine learning (ML) models. This is particularly important in the age of data-hungry deep learning models, where black-box models have repeatedly been demonstrated to overfit spurious patterns despite the abundance of training data. Relatedly, I think explainability, along with other inductive biases such as structured representations (*e.g.*, syntactic parses of sentences [4, 7, 6, 16]), play a crucial role in making sure AI systems generalize beyond the training distribution. Thus, I am also deeply interested in enabling explainable compositional reasoning and incorporation of declarative knowledge in powerful statistical learning systems through more computationally efficient approaches than data augmentation.

In the short term, I would like to focus on two main directions of investigation. One of them is continued research into multi-turn information-serving agents [15], where data scarcity is likely to remain an issue in the foreseeable future, requiring more explainable and data-efficient approaches. Moreover, I believe solving this important NLP problem can also lead to novel ML techniques and evaluation methods that are applicable to other (even non-NLP) tasks that involve long trajectories of agent exploration. The other direction is at the intersection between complex reasoning and model compositionality. On the one hand, I would like to investigate how to better leverage existing structured information (e.g., parse trees) to guide representation learning and compositional reasoning in an explainable manner. On the other hand, I think it is also important that we develop better structured prediction and compositional models, both task-agnostic ones that generalize, and task-specific ones that incorporate task insights as inductive bias.

Overall, I hope to demonstrate in my research, through tackling important NLP and ML problems, that explainable models can not only enable us to solve tasks that are otherwise impossible without huge amounts of data, but also be more robust to adversarial examples and have more predictable worst-case behavior.