Research Statement

Harsh Jhamtani
Carnegie Mellon University

The interface between human users and large collections of data is an important application area for artificial intelligence (AI) technologies. But this represents a challenge - how can we build systems that reason about data and present their results effectively in natural language? Can we leverage natural language in order to help systems learn useful abstractions of data similar to what humans often do? Can we learn which effectively identify salient or important patterns? Can we do all this in ways which build user trust in machine learning models?

A. Describing salient trends in tabular data based on inferences on multiple cells

B. Describing salient differences between two similar images while screening out noisy uninteresting aspects (e.g. some leaves moving in background)

Figure 1

Much recent progress in various data-to-text tasks has relied on deep learning, often using neural networks with soft attention mechanisms to select salient aspects from data, and constructing fluent natural language text. However, in naturally occurring descriptions of data, humans often refer to higher-level patterns which may require complex computations on data. Consider a numerical table consisting of stock prices of a company over a period of time. A simple description would require simply selecting some value and state it in natural language. In contrast to it, a more complex pattern, such as 'Stock for DJI rose more steeply over last week compared to stock GSPC', will require much higher level reasoning over multiple cells/values (**Figure 1(A)**). But how do we effectively learn valid and interesting aggregations of values. Can we do so in a manner which also exposes some abstractions of the underlying computations and reasoning leading to such descriptions? Turns out that such higher level patterns often cannot be extracted using neural models with attention alone. Moreover, even for cases where it can model some of the complex patterns, such models tend to imprecise (e.g. generate hallucinated descriptions), and offer little insights to a user about its working.

As another example, consider the task of describing differences between two similar images ('spot-the-diff'). An example is shown in **Figure 1(B)**. Of course, individual pixels will vary across images due to minor changes in lighting and even sensor noise. However, these differences are not typically salient to humans and would be screened out by a person performing the task. Moreover, humans tend to group together related pixel changes and even combine together related changes, abstracting to higher-level descriptions. For example, if two people both move across the scene, this change is likely to be described in terms of a 'pair' rather than two separate changes. Can we design models which explicitly incorporate and expose such hierarchical reasoning structures (pixel-level filtering, grouping related changes, inferring important from unimportant changes)? Do such models work better than contemporary approaches?

My research focuses on inducing or incorporating such useful (and often multi-level) discrete structures on data as well as output natural language text. Throughout my work I used a variety of discrete structures based on task suitability, including but not limited to sequence of deterministic programs of increasingly abstract nature, latent alignments between portions of data and text, sparse hierarchical grouping, latent reasoning chains on knowledge bases, and so on. For a more concrete example, lets revisit the numerical time series data captioning. A relatively straightforward way to model this would be apply a neural encoder on the time series data – however, as we observe in our experiments, such models fail to capture more complex patterns, and often generate incorrect outputs. In contrast, we propose methods to induce separate modules to detect useful trends such as peak or dip in time series numerical data, being guided only by accompanying natural language descriptions. The modules represent understandable and useful abstractions, and can be combined to form more complex patterns which go beyond value selection. The proposed model exposes an interpretable computation structure, and leads to very high precision in output text. Such is the value of incorporating useful structures for uncovering complex latent patterns in data to text problems.

The rest of the research statement is divided into two parts. In the first part, I discuss incorporating certain hierarchical discrete structures in model architectures for different grounded natural language generation tasks such as image and table captioning, dialog response generation, and reasoning chains for question-answering – spanning across various data modalities (images, tabular data, numerical data, knowledge bases). In the second part, I discuss methods that use discrete structures to induce latent global patterns in text generation, such as rhyming patterns in poetry and generation plans in narrative text

generation. I will discuss when are such structures useful, and how the resulting models can better characterize complex patterns on data, tend to be less data hungry, and more interpretable compared to many contemporary methods relying on neural attention alone. Additionally, I will briefly discuss how training such models maybe challenging, and propose some novel training methods for the same.

1 Inducing Salient Abstractions for Human-like Data Descriptions

1.1 Modeling Multiple Levels of Visual Salience for Difference Descriptions

In [1], I introduced a novel task of learning to describe all the differences between two similar images ('spot-the-diff'). An example is shown in Figure 2(A). Solving such a task has applications in assisted home monitoring, tracking of enterprise digital assets, and tracking and understanding fake art.

As mentioned earlier, solving such a task requires modeling of perceptual salience structures, and identifying appropriate level of abstractions. Individual pixels will vary across images due to minor changes in lighting and even sensor noise. However, humans are typically interested in more meaningful and salient differences. Moreover, a model for the task should not overwhelm the user with each individual difference – combine together related changes, abstracting to higher-level descriptions. For example, if two people both move across the scene, this change is likely to be described in terms of a 'pair' rather than two separate changes. We propose models which meaningfully group together differing pixels in order to approximate object-level differences. We use discrete latent variables to align clusters of differing pixels with output sentences - the latent variables being unobserved during training. Our joint learning procedure induces a model of salience without direct supervision, being guided by the accompanying human written text only.

Difference description generation can have various applications in non-image data domain as well. For example, I worked on generating chess commentary for a game move [7] – essentially viewing it as identifying the salient changes in game state. We propose models for a chess move commentary generation which work by comparing successive chess board states through hierarchy of increasingly more abstract features: piece movement, piece interactions, and overall game score changes.

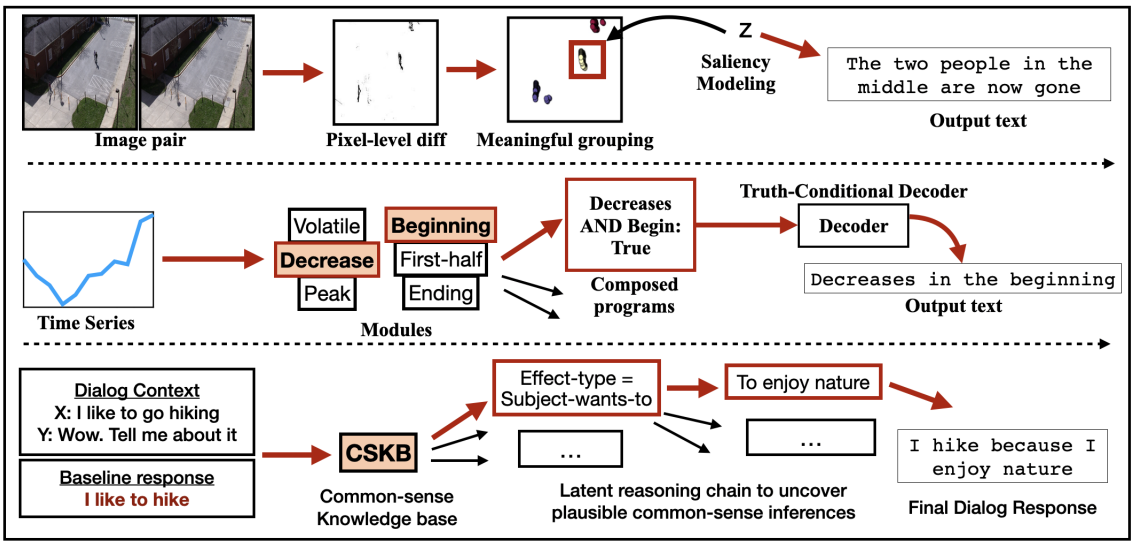

Figure 2: Inducing salient abstractions on data via interpretable hierarchical operations.
A: Spot-the-diff: Aim is to describe differences between pairs of similar images. We propose a model which identifies salient changes at multiple abstractions – (pixel level filtering → object level grouping → latent discrete alignment).
B: Time Series Captioning: We propose a model which induces hierarchy of composable useful operations (convolutional operations → modules → programs → valid programs).
C: Open-domain dialog response generation: We propose a model which can can identify plausible common-sense reasoning chain inferences for a given bot-persona attribute or a default response from background dialog model, and leverages it to construct to more interesting/engaging responses.

1.2 Inducing Neuro-Symbolic Patterns for Numerical Data

Now lets consider the task of identifying and describing interesting patterns in tabular data. Such a technology has many applications such as a digital assistant for spreadsheet analysis. Identifying candidate patterns (such as a dip) would involve identifying which mathematical operations (such as subtraction or slope) to execute on (multiple) cells, and learning to identify salient patterns (from accompanying natural language descriptions in training data).

In [4], we propose methods to induce modules which detect useful trends such as peak or dip in time series numerical data, being guided only by accompanying natural language descriptions. The modules are combined via a latent computation graph to form programs, which outputs truth value of whether the feature represented by the composed computation graph holds true for a given data point or not (Figure 2(B)). The resulting model gives highly precise outputs based on learned salient patterns. Additionally, due to the modular nature of the latent programs, we observe that model can generalize very well on unseen (at training) combinations of modules, and tail of the distribution of patterns. Such ideas have applications in other domains as well – In [9], we propose techniques for inducing sparse word embedding vectors guided by topics, leading to more interpretable word embeddings without decrease in downstream performance.

1.3 Fine-grained Reasoning for Knowledge Base Grounded Text Generation

Often lot of useful structured and unstructured information is stored in knowledge base stores. There has been a long line of work in leveraging such information by machine learning models for generating useful outputs. A popular contemporary approach for KB-grounded generation is to encode a large dump of retrieved (somewhat) relevant information from the knowledge base. However, often all information is not useful – extraction of useful bits of information from knowledge base for a given context often requires a principled way of reasoning through it.

We propose models which induce latent reasoning structures for extracting useful information from knowledge bases. Often such reasoning structures involve fine grained selection of sentences from knowledge base, as well as carefully combining the relevant pieces. Additionally, our models expose the underlying reasoning structures which can inspire trust from users.

In [11] we propose and discuss models for persona grounded dialog, which can leverage an external common-sense knowledge base based on given persona, and perform fine-grained selection (unobserved in data) of persona attribute (Figure 2(C)). In [5], we discuss uncovering abstract reasoning chains as explanations for open domain multi-hop reasoning questions. Compared to prior work which uses a dump of relevant sentences, we focus on selection and combinations of sentences which form valid reasoning chains. In both the cases we observe that the proposed models lead to improved generalization compared to baselines, and uncover interesting reasoning patterns.

2 Discrete Plans for Long-form Text Generation

Till now, I discussed incorporating discrete latent structures on data. I will now discuss how related techniques can be leveraged to induce latent global plans for text generation. Similar to data, multi-sentence text often demonstrates complex structures. Many neural language models often fail to capture such high-level structures present in text: for example, rhyming patterns present in poetry or narrative plans for coherent long form text generation. Prior work has heavily relied on injection of external knowledge such as rhyming knowledge, or use external tools to tag narrative plans, to effectively model such long range patterns. I focus on investigating whether such long range patterns or structures can be treated as latent variables, and be learned from the data, resulting in better quality outputs.

2.1 Latent Discrete Generation Plans for Coherent Narrative Generation

Automatic and/or machine-assisted text generation has recently gained lots of interests. Given a rough sketch of piece of text, such as using keywords, can machines help people in authoring new content? A logical and coherent structure is an important characteristic of easily comprehensible text. Towards this end, prior work has attempted to learn text generators conditioned on an outline of keywords. However, naturally occurring data is not tagged with such plans, and therefore prior work has relied on heuristic tagging of plans.

In [3], we propose a deep generative model which uses a discrete latent plan realized via a sequence of keywords, which loosely determine the story plot. Such a model, however, cannot be trained by maximizing the likelihood due the hierarchical latent variables. We propose methods to effectively train such a model using variational learning to approximate a lower bound on the likelihood. Going forward, I am looking to extend this work to use more expressive latent plans, and exploring how to use similar techniques with decoders initialized using large pre-trained language models.

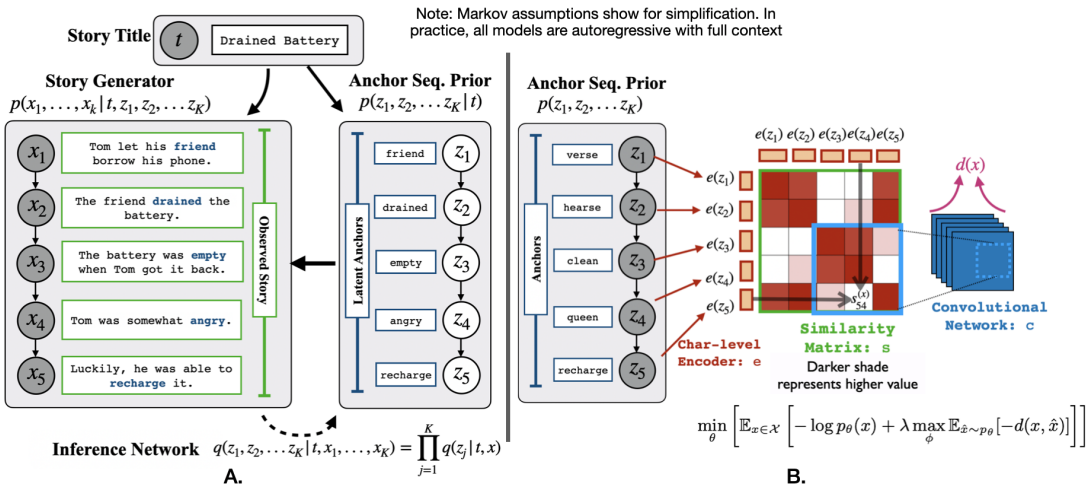

Figure 3: Discrete Generation Plans for Long-form Text Generation. We propose models which induce useful discrete plans for multi-sentence text generation.
A: Latent Discrete Plans for Coherent Narrative Generation: For the task of narrative generation, we propose a deep latent variable model which induces latent plots (sequences of keywords) loosely determining the narrative structure.
B: Structured Discriminators: For more formal structures such as rhyming schemes, we propose a novel training scheme by incorporating structure into an adversary. More specifically, our discriminator/adversary operates only on pairwise cosine similarities of learned word representations. Discriminator induces a rhyming metric, while generator learns plans which follow rhyming schemes present in data.

2.2 Structured Discriminators for Modeling Long-Range Latent Patterns

While the previously mentioned model is applicable for coherent paragraph generation, the notion of plan was loosely defined via a sequence of keywords which can guide in generating coherent stories. However, in some cases, we need to model much more formal structures in language such as rhyming schemes.

In [8], we propose a novel generative adversarial setup where we insert inductive bias into the discriminator. More specifically, instead of using a monolithic discriminator, we choose its architecture carefully. We design the discriminator to focus on long-range structural properties of its input. The proposed structured discriminator learns to identify rhyming word pairs as well as rhyming constraints present in the poetry datasets without being provided phonetic information in advance. This idea can be extended to learning other long-range patterns in text, and has applicability even beyond text data. In fact, in [2], we extend the above idea for music generation focusing on self-repetition to learn a model which can generate compositions with long range repetition patterns.

3 Summary

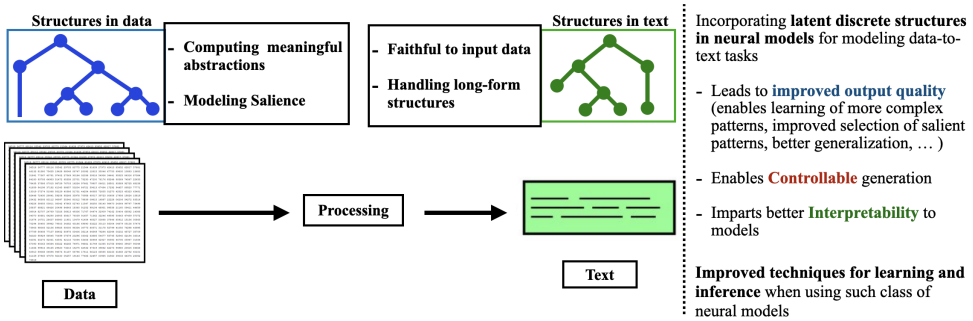

Figure 4: Research Summary: My research focuses on developing models that induces meaningful latent discrete structures for grounded natural language generation. Compared to using attention alone, such structures can often better model complex patterns on data, tend to be less data hungry, enable controllable generation, and more interpretable and robust compared to many contemporary methods relying on neural attention alone.

My research focus has been on uncovering latent discrete structures on data as well as output text for Grounded Natural

Language Generation (Figure 4). As I show through specific instantiations of the general idea, the resulting neural models often demonstrate improved text quality (via handling of complex high-level descriptions, interpretable selection of salient patterns in data, preciseness/faithfulness of output text to input data, improved robustness, etc.) as per various automated metrics and human evaluations. The resulting models expose useful abstractions, which often enabling controllable and interpretable generation.

Future Directions: Much of the work discussed in the statement deals with static datasets and static models. More challenging situations may encompass evolving datasets, and need to update models based on new information, data, or even feedback from users. Going ahead, I am interested in exploring how the modeling/enforcing of structures can help in data-to-text tasks with dynamic data sources. Towards this direction, I have done some initial explorations on how to incorporate knowledge sources for dialog response generation using gradient-based decoding without the need for re-training the models [10, 6].

References

[1] JHAMTANI, H., AND BERG-KIRKPATRICK, T. Learning to describe differences between pairs of similar images. In *EMNLP 2018* (2018).

[2] JHAMTANI, H., AND BERG-KIRKPATRICK, T. Modeling self-repetition in music generation using generative adversarial networks. *Machine Learning for Music Discovery Workshop, ICML* (2019).

[3] JHAMTANI, H., AND BERG-KIRKPATRICK, T. Narrative text generation with a latent discrete plan. In *Findings of EMNLP 2020* (2020).

[4] JHAMTANI, H., AND BERG-KIRKPATRICK, T. Truth conditional time series description. In *Under Review* (2021).

[5] JHAMTANI, H., AND CLARK, P. Learning to explain: Datasets and models for identifying valid reasoning chains in multihop question-answering. In *EMNLP* (2020).

[6] JHAMTANI*, H., GANGAL*, V., HOVY, E., AND BERG-KIRKPATRICK, T. Improving automated evaluation of open domain dialog via diverse reference augmentation. In *Findings of ACL* (2021).

[7] JHAMTANI*, H., GANGAL*, V., HOVY, E., NEUBIG, G., AND BERG-KIRKPATRICK, T. Learning to generate move-by-move commentary for chess games from large-scale social forum data. In *ACL* (2018).

[8] JHAMTANI, H., MEHTA, S. V., CARBONELL, J., AND BERG-KIRKPATRICK, T. Learning rhyming constraints using structured adversaries. In *EMNLP 2019* (November 2019).

[9] JHAMTANI, H., SUBRAMANIAN, A., PRUTHI, D., BERG-KIRKPATRICK, T., AND HOVY, E. Spine: Sparse interpretable neural embeddings. In *AAAI-18* (2018).

[10] MAJUMDER, B. P., BERG-KIRKPATRICK, T., MCAULEY, J., AND JHAMTANI, H. Unsupervised enrichment of persona-grounded dialog with background stories. In *ACL* (2021).

[11] MAJUMDER, B. P., JHAMTANI, H., BERG-KIRKPATRICK, T., AND MCAULEY, J. Like hiking? you probably enjoy nature: Persona-grounded dialog with commonsense expansions. In *EMNLP 2020* (2020).