RESEARCH INTERESTS

Guanlin Li
Department of Computer Science
Harbin Institute of Technology
Harbin, China
{epsilonlee.green}@gmail.com

> *"You go to a great school not so much for knowledge as for arts and habits; for the habit of attention, for the art of expression, for the art of assuming, at a moment's notice, a new intellectual position,* ***for the art of entering quickly into another person's thoughts****, for the habit of submitting to censure and refutation, for the art of indicating assent or dissent in graduated terms... And above all, you go to a great school for self-knowledge."*
>
> — William Johnson Corg

1 PRELUDE - A HISTORICAL VIEW OF MYSELF

I began my journey in Natural Language Processing (NLP) from an ambition of making human-like Natural Language Generation (NLG) agent who can consciously generate language intents or speech acts to instantiate with utterances (Reiter & Dale, 2000). After digging deeply into the logic, semantic representation side of macro and micro planning of an NLG system. I quickly find that human as its designer should come up with a large amount of rules for making the system actually work. This system engineering view of NLG can make the final deployed system very useful and accurate in specific domain, and also trustworthy since human have controlled every single rule of its inner mechanism. However, this is far from what I initially believe into as mentioned above. The consciousness of an NLG system can never be achieved. Then, I turn to the data-driven paradigm dubbed as learning, which I found very appealing, since the agent can use the experience (labelled data mostly) to obtain the strategy for solving similar but unseen new problems. However I later on I found that open-domain dialogue system as an NLG agent is notoriously hard to design and the task is *extremely* ill-posed in essence [1]. So I began to accept the current hardness and impossibility of building conscious NLG agent, and begin to embrace the central idea of machine learning - **generalization** (Zhang et al., 2017). Once again I turn to the study of machine translation which has better evaluations and well-defined, since given the input, the output can be said to be determined at least in its semantic meaning. And the *word alignment* as an intuitive surface form evidence of translation functions as the pillar philosophy of formulating both statistical (Koehn, 2010) and neural machine translation (Bahdanau et al., 2015). The deterministic aspect of the task makes it an excellent venue for studying realistic generalization of learning machines.

2 RESEARCH INTERESTS

I broadly identify with the community of *reliable* Artificial Intelligence (AI) and those researches on both theoretical and empirical aspects of current data-driven AI paradigm or **machine learning**. I think just in the near future (maybe within ten years), it is of great difficulty in building one *holistic* agent embodied with every possible abilities of human intelligence, though I deeply appreciate the ongoing **BabyAI** project at MILA. They are researching along the frontiers of building human-like agent that can perceive *multi-modal* input, which might make breakthrough on the **representation** aspect of AI [2]. However, during the process, I think a better compromise is to design reliable AI systems that are *trustworthy* for human users and adapts according to various security guarantees. This can democratize AI and make AI accessible to all ordinary people.

[1] In my opinion, I see a prediction problem as ill-posed if the conditional entropy $\mathbb{H}(Y|X)$ is very high, which means there are many possible solutions among which exists few statistical connections.

[2] Mentioned in John Langford's Machine Learning the Future lectures.

Adversarial examples are pervasive monsters for all kinds of machine learning models. Currently we have *not* yet obtained widely accepted theoretical insights into *why* they exist and their connection to the general generalization ability of model. Some works argue that adversarials are an aspect of out-of-distribution generalization, which we should turn to study in depth instead (Yin et al., 2019). I think the study of adversarial examples can shed some lights on the study of **realistic generalization** and *vice versa* if we could design an approximate distributional model of the underlying task at hand with current deep generative models (Brown et al., 2020) with strong *interpolation* ability.

In terms of model's generalization ability, researches on non-vacuous *data-dependent* generalization bounds for deep neural networks (Dziugaite et al., 2020) motivates myself to put an eye on the value of training data (the experience for the model to learn from). That is, given a *limited sized* training set $\mathcal{D}_{tr}$ for a real-world task $\mathcal{T}$, we *cannot* expect the model to generalize well on instance $x \notin \mathcal{D}_{tr}$ that requires *"knowledge"* which is actually out of the scope of $\mathcal{D}_{tr}$. Therefore, we should identify the limitations of the generalization ability of a model, so as to provide reliable performance guarantee for task $\mathcal{T}$. Universally applicable solutions include out-of-distribution detection or anomaly detection (Ruff et al., 2020), which we can leverage as a security door for any task before sending the input to the prediction model. A more advanced take might be designing a built-in mechanism such as certified guarantee for adversarial robustness (Raghunathan et al., 2018). This requires us to make assumptions, which is model-agnostic and $\mathcal{D}_{tr}$-dependent, about task $\mathcal{T}$, To be honest, the above mentioned solutions all demand an operational or theoretical understanding of *what* a model rely on for generalization on x with which kinds of properties (Nagarajan et al., 2020).

Besides all, since I have mainly done researches on machine translation pursuing my doctor's degree, I develop my sense that every realistic task has its own **structure** (for example, word or phrase alignment as *axiomatic* translation knowledge the model can rely upon) and special property which the general statistical learning theory fails to take into account. So to develop theory for realistic generalization on certain task $\mathcal{T}$ requires better understanding the task itself. This may starts from uncovering and understanding every failure modes of the model's generalization ability and making exposed learning and generalization phenomenon more transparent and its connection to other learning paradigms such as *transfer learning*, *domain adaptation*, *semi-supervised learning* and *self-supervised learning*. I see these mentioned learning paradigms as a holistic solution to improve the model's generalization ability from irrelevant or unlabeled data beyond $\mathcal{D}_{tr}$.

Based on the above humble discussion, I am extraordinarily interested in the following topics.

- *(Realistic)* **generalization** and **compositionality**
- **Interpretability** and **explanability**
- **Machine translation** and **multilinguality**

In the following subsections, I will make concise discussions about each of these topics and develop my very focus about some potential roads for approaching these exciting directions.

2.1 Realistic generalization and compositionality

Formal language Formal languages are largely simplified versions of natural language, and we can give precise description of them. After attending the workshop on **Deep Learning and Formal Languages** at ACL 2019, I got the feeling that if we can largely constrain the abundant linguistic phenomenon [3] in natural language to construct a delicate approximate artificial language, we can learn from the learning effects on the formal language and transfer the understandings to handle challenges in the realistic setting. This approach actually matches the philosophy of the *simulation* and *sim2real* approach in the self-driving cars and robotic manipulation literature (Weng, 2019).

Compositionality Compositionality is a concept and principle from philosophy of linguistics which advocates that meaning (semantics) is constructed through axiomatic meanings with a system of composition (syntax). From a compositional perspective, we can understand the generealization of a machine translation model on an input $x \notin \mathcal{D}_{tr}$ as entailed by its generalization on every natural and meaningful sub-sequences in x. Or we can find the most relevant training instances that are responsible in a compositional way to the generalization on x. This can reveal the limitation of $\mathcal{D}_{tr}$.

[3] This might require detection and identification of those task-specific **long-tailed** linguistic phenomenon, for instances, low-frequency lexicons, strange syntactic variations and construction-like utterances etc.

2.2 Interpretability and Explanability

I think of interpretability as a way of explaining why the model makes certain prediction to end user, and if the features the model relies upon for making prediction are intuitive and can be understood by the end users, the goal of interpretability is achieved. This might further influence the user's trust on using the model. However, many features in deep models cannot make sense to end users, so one possible solution is to repurpose the uninterpretable features into interpretable ones through certain transformation e.g. clustering and then annotated by human with interpretable *concepts*. This have been done in image recognition tasks (Koh et al., 2020).

2.3 Machine translation and multilinguality

As I mentioned before, how to leverage abundant unsupervised data for augmenting the "knowledge" that does not exist or learnable by the model from $\mathcal{D}_{tr}$ is the current trends of NLP community. The success of pretraining, in my humble opinion, seems to come from a better similarity relationship between different words and their composed phrases is established. This makes over-fitting in low-data regime less severe to influence generalization. One interesting question is can unsupervised data substitute supervised data in $\mathcal{D}_{tr}$? If can, to what extent?

Knowledge removal One interesting direction is to remove certain knowledge in $\mathcal{D}_tr$, for example, removing all possible alignment knowledge of *(bank, banco)*. Can unsupervised data help the model learn such alignment? If can, what kind of unsupervised data it should look like?

Dataset reduction Motivated by the work of dataset distillation (Wang et al., 2018), to what extent we can reduce the size of $\mathcal{D}_{tr}$ and use unsupervised data to preserve the original performance. This can somehow shed light on the corset selection problem of $\mathcal{D}_{tr}$ as well as the degree of information or knowledge redundancy in $\mathcal{D}_{tr}$. As for the task of neural machine translation where inference is an approximate search algorithm, the dataset reduction approach can also shed light on how much redundancy of word alignment knowledge can guarantee the search behavior to predict the correct word translation?

Multilinguality Since the model capacity becomes larger and larger, it can memorize and learn as many as possible languages at one time. A straight forward question here is that: does the learning from the hybrid corpus result in so-called universal representations among different language? If so, how does it emerge during alignment-unaware training? What does the universality embody? The semantic meaning or some word order (syntax pattern)? How the universality or specificity of the representation varies in terms of linguistic topology? Can disentanglement of them lead to controllable transferibility between different languages?

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2015.

T. Brown, B. Mann, Nick Ryder, Melanie Subbiah, J. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, G. Krüger, Tom Henighan, R. Child, Aditya Ramesh, D. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, E. Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, J. Clark, Christopher Berner, Sam McCandlish, A. Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020.

G. Dziugaite, Kyle Hsu, Waseem Gharbieh, and D. Roy. On the role of data in pac-bayes bounds. *ArXiv*, abs/2006.10929, 2020.

Philipp Koehn. *Statistical Machine Translation*. Cambridge University Press, USA, 1st edition, 2010. ISBN 0521874157.

Pang Wei Koh, T. Nguyen, Yew Siang Tang, Stephen Mussmann, E. Pierson, Been Kim, and Percy Liang. Concept bottleneck models. *ArXiv*, abs/2007.04612, 2020.

Vaishnavh Nagarajan, A. Andreassen, and Behnam Neyshabur. Understanding the failure modes of out-of-distribution generalization. *ArXiv*, abs/2010.15775, 2020.

Aditi Raghunathan, J. Steinhardt, and Percy Liang. Certified defenses against adversarial examples. *ArXiv*, abs/1801.09344, 2018.

Ehud Reiter and Robert Dale. *Building Natural Language Generation Systems*. Cambridge University Press, USA, 2000. ISBN 0521620368.

Lukas Ruff, J. Kauffmann, Robert A. Vandermeulen, Grégoire Montavon, W. Samek, Marius Kloft, Thomas G. Dietterich, and K. Muller. A unifying review of deep and shallow anomaly detection. *ArXiv*, abs/2009.11732, 2020.

Tongzhou Wang, Jun-Yan Zhu, A. Torralba, and Alexei A. Efros. Dataset distillation. *ArXiv*, abs/1811.10959, 2018.

Lilian Weng. Domain randomization for sim2real transfer. *lilianweng.github.io/lil-log*, 2019. URL `http://lilianweng.github.io/lil-log/2019/05/04/domain-randomization.html`.

Dong Yin, Raphael Gontijo Lopes, Jonathon Shlens, E. Cubuk, and J. Gilmer. A fourier perspective on model robustness in computer vision. *ArXiv*, abs/1906.08988, 2019.

C. Zhang, S. Bengio, M. Hardt, B. Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *ArXiv*, abs/1611.03530, 2017.