Constrained Inference for Bridging the Distributional Gap in Natural Language Processing

TAO MENG, DEPARTMENT OF COMPUTER SCIENCE, UCLA

Machine learning techniques have achieved remarkable performance in a variety of tasks. However, usually they are evaluated on data from the same distribution as training. When we deploy the machine learning model to real-world application, it is common that the data we do prediction on has a different distribution from the training data. We call this kind of distribution difference as distribution gaps. Distribution gaps could come from the application setting like transfer learning, or from the dataset bias existing in data collection process [3], causing performance drop on the out-of-distribution data in real-world application.

To compete against the gaps, I purpose to leverage constraints, a set of specific rules that the model requires to follow in instances or distributions, in machine learning models. My research goal is to bridge the distribution gaps via constraints. Specifically, I aim to 1) compile human knowledge into constraints and inject them into machine learning models to boost the out-of-domain performance; 2) design a constraints learning framework to automatically detect the distribution gaps in forms of constraints. Generally, my research is bridging the model performance between training and test distribution, and bridging human knowledge to neural models.

1 Research Interests and Prior Research Achievements

My vision is supported by my past research achievements in injecting constraints in natural language processing (NLP) applications, and the successful attempt on building a framework for learning linear constraints from data.

Cross-lingual Dependency Parsing with Word Order Constraints [7] In this work we leverage constraints to bridge the gap between the source and target language in transfer learning. Dependency parsing is a classical NLP task to analyze the grammatical structure of a sentence. Neural models perform well on rich-resource languages like English but fails in those low-resource languages. Thus people train on those rich-resource languages called source languages, and apply them on the target languages. In this process, models leverage some language-dependent features, in particular, word order features, to make decisions. We purpose corpus-level constraints on word order features and corresponding inference algorithms, and compile linguistic knowledge for different languages into constraints. We show that we improve the model performance on different target languages by bridging the source-target distribution gap with the corresponding constraints.

Mitigating Gender Bias Amplification in Visual Semantic Role Labeling (vSRL) [4] In this work we show that benchmark dataset is collected with gender bias, causing a distributional gap from real world. Specifically, vSRL is a task that given an image with a human, we predict the activity in the image and corresponding attributes of the human including gender. The images are biased in gender about some activities, e.g., in driving images there are more males than females. Models trained on a it even amplify the bias, which is potentially risky in causing society issues [9]. With our designed constraints on gender feature, we are able to mitigate this amplification

behaviour by regularizing the posterior distribution without hurting the model performance, hence avoid the society issues when we deploy the model in real world applications.

Integer Linear Programming (ILP) Framework for Constraints Learning [5] In this work we purpose an ILP framework to learn linear constraints from data. In my previous work constraints are pre-defined by human. Sometimes constraints can be implicit or the number of constraints can be large. Thus, we aim to build a framework to automatically learn constraints from data, and are able to smoothly incorporate with deep neural networks. We show that we are able to learn the constraints formulating the structure of the label space, which can be challenging to identify by neural architectures. The constraints incorporated model achieve better performance with better understanding about the task.

Controllable Text Generation with Neurally-Decomposed Oracle [6] In this work, we propose a general and efficient framework to control auto-regressive generation models with NeurAlly-Decomposed Oracle (NADO). Given a pre-trained base language model and a sequence-level boolean oracle function, we propose to decompose the oracle function into token-level guidance to steer the base model in text generation. Specifically, the token-level guidance is approximated by a neural model trained with examples sampled from the base model, demanding no additional auxiliary labeled data. We present the closed-form optimal solution to incorporate the token-level guidance into the base model for controllable generation. We further provide a theoretical analysis of how the approximation quality of NADO affects the controllable generation results. Experiments conducted on two applications: (1) text generation with lexical constraints and (2) machine translation with formality control demonstrate that our framework efficiently guides the base model towards the given oracle while maintaining high generation quality.

2 Future Research

My long-term research goal is to design a flexible machine learning framework that models are capable to adapt to different application distributions by injecting human knowledge and constraints learned automatically. Basically I would like to combine the constraints incorporation direction and the constraints learning direction together. This whole framework can be further applied on broader applications. To achieve this, some concrete directions are listed below:

- **Controllable Generation on Large Pretrained Language Models (LLMs)** LLMs such as GPT-3 [1], PaLM [2] or OPT[8] have got remarkable achievements on complicated tasks. Those LLMs can be treated as large knowledge base and people design task-specific prompts to query the models. Considering my success on controllable generation, I plan to seek for the possibility that querying the LLMs by controllable generation. Specifically, we can choose proper oracle or constraint for the specific task, and control the LLMs distribution to focus on the given task. In my prior work we demonstrate that the results given by the controlled model keep the generation quality as well as following the controlling criteria. In this project I hope we can find an efficient approach to take advantage of good geneartion quality of LLMs while controlling it to work on the specific task.

- **Automatically Mining Constraints about Distribution Gap.** It would be helpful if we can detect and formulate the distribution gap given two distribution efficiently, and compile them into constraints. This work benefits machine learning research mainly in the following two scenarios: 1) For transfer learning we will be able to formulate the difference between the source and target. We can design better transfer strategies with such information then. 2) For the training data containing spurious features or dataset bias, we can detect them given some real-world samples. Based on this we can do debug to the model and corresponding training dataset.

- **Combination of Constraints Mining and Constrained Inference.** The constraints we mine from data will finally be used for helping the model. Thus, one of my research goal is to combine them together. Given two distributions, we automatically mine constraints to bridge the distribution gaps and use the constraints to guide the model do adaptive predictions on them. This research allows me to connect my work together to make a complete end-to-end framework.

[1] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In *NeurIPS*, 2020.

[2] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, and N. Fiedel. Palm: Scaling language modeling with pathways. *CoRR*, abs/2204.02311, 2022.

[3] S. Gururangan, S. Swayamdipta, O. Levy, R. Schwartz, S. R. Bowman, and N. A. Smith. Annotation artifacts in natural language inference data. In *NAACL-HLT (2)*, pages 107–112. Association for Computational Linguistics, 2018.

[4] S. Jia, T. Meng, J. Zhao, and K. Chang. Mitigating gender bias amplification in distribution by posterior regularization. In *ACL*. Association for Computational Linguistics, 2020.

[5] T. Meng and K. Chang. An integer linear programming framework for mining constraints from data. In *ICML*, 2021.

[6] T. Meng, S. Lu, N. Peng, and K.-W. Chang. Controllable text generation with neurally-decomposed oracle. In *NeurIPS*, 2022.

[7] T. Meng, N. Peng, and K. Chang. Target language-aware constrained inference for cross-lingual dependency parsing. In *EMNLP/IJCNLP (1)*. Association for Computational Linguistics, 2019.

[8] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. T. Diab, X. Li, X. V. Lin, T. Mihaylov, M. Ott, S. Shleifer, K. Shuster, D. Simig, P. S. Koura, A. Sridhar, T. Wang, and L. Zettlemoyer. OPT: open pre-trained transformer language models. *CoRR*, abs/2205.01068, 2022.

[9] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K. Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *EMNLP*, 2017.