Research Statement – Yue Dong

McGill University/Mila, January 2022

Natural language processing (NLP) aims to understand and process human language by machines. Many NLP tasks such as sentiment analysis and reading comprehension have come of age thanks to advances in machine learning. For example, machine reading comprehension models have exceeded human performance on both the exact match (EM) and F1 score on the task of *question answering* (i.e. SQuAD 2.0 dataset). NLP has reached a level of maturity long-awaited by people to use them in practical settings, for example, smart assistants like Apple Siri or Amazon Alexa that help people inform complex decisions quickly. By utilizing billions of documents across many domains, pre-trained neural models have become crucial building blocks to other higher-level natural language generation (NLG) tasks that our field has yet to accomplish for practical employment, such as text summarization and human-level dialog.

In order for machine learning systems to be effectively and broadly used in practical settings, **we require models that are both robust for generalization and reliable with faithful information.** Current neural models have achieved state-of-the-art (SOTA) performances on benchmark datasets in limited domains, where expensive annotated data are available (e.g. news). These models are typically neither trained nor evaluated for performance out-of-domain. Besides, despite models having SOTA performances in data-rich settings, they still suffer from content hallucination (Reiter, 2018). These hallucinations, which are either irrelevant or contradicted to the inputs, have become a notable obstacle to the adoption of neural models as practical tools. Users also have **practical concerns about model biases, fairness, and ethics violations.** Despite sophisticated neural architectures that can exceed human performances on benchmark datasets, employing them in real-life settings still remains challenging due to the aforementioned concerns.

My research addresses fundamental challenges faced when leveraging natural language generation techniques on far-reaching practical problems. Towards this end, I provide fundamental advances in domain robustness and generation faithfulness. To **facilitate robust generalization** I design new training and inference algorithms cognizant of strengths in the reinforcement learning algorithm (Dong, Shen, et al., 2018), loss regularization (Dong*, Grenander*, et al., 2019), and attention mechanism (Dong, Bhagavatula, et al., 2021), and eliminate redundant computation through jointly multi-tasking learning (P. Liu, Dong*, et al., 2019). To support generalization to **data-scarce domains** (e.g. scientific and medical texts), I incorporate domain-specific discourse information (Dong, Mircea, et al., 2021) to graph-based models to improve generalization, in addition to creating several datasets in the scientific domain in multi-document setting (Lu, Dong, et al., 2020) and single-document setting (Meng, Thaker, et al., 2021). To **advance faithful text generation**, I design edit-based models for facts preserving (Dong, Li, et al., 2019), and utilize adversarial training (Cao, Dong, et al., 2020) and implicit knowledge (Dong, Wang, et al., 2020) for hallucination reduction. I have also worked on the practical problem of anomaly detection in streaming data (Dong & Japkowicz, 2016, 2018), where the idea of using neural networks, instead of the popular choices of decision trees at the time, has won the **best paper award** in Canadian AI conference. In conjunction with these new machine learning techniques, I will collaborate with domain experts to make a positive mark on society (§4).

1 Balancing content learning and structural bias in NLG

Neural-based models that are trained on large-scale text datasets often automatically capture patterns and biases in natural language. These biases can be divided into two categories: 1) structural biases that reflect human writing patterns for different genres of texts; and 2) linguistic regularity biases that mirror human biases (e.g. racism). My past work mainly focuses on structural biases, which provide a shortcut for neural models from learning the actual semantic contents. For example, in news summarization, selecting the first few sentences as a summary (i.e. lead baseline) is hard to beat, as journalists are instructed to follow the pyramid scheme and write the most important information in the opening sentences of a news article (Brandow, Mitze, et al., 1995; Nenkova, 2005). This lead bias is unfortunately overlooked in many sequential extractive summarization models to the extend of trading off content learning.

I have developed bias-conscious techniques that balance content learning and structural bias in text summarization. In particular, I proposed to process the extractive summarization task with a non-autoregressive model (i.e. a multi-armed contextual bandit) that is trained with a policy-gradient reinforcement learning algorithm (Dong, Shen, et al., 2018). This non-autoregressive setting allows our model to bypass the inherent lead bias of the sequential labeling approaches, where the model tends to select early sentences from the documents. I showed that by **designing a framework that does not inherently have structural biases**, we can generalize into sub-datasets where good summary sentences appear late in the document by **a performance improvement up to 35%**. Besides, with collaborators we further explore the sentence representation learning with non-autoregressive models by imposing an auxiliary loss in REINFORCE (Williams, 1992) that drives the model to focus on content learning (Dong*, Grenander*, et al., 2019).

2 Generalizing to new natural language domains

Generalization to out-of-domain data is an ability natural to humans yet challenging for machines. We need models that generalize to diverse texts from different domains without expensive re-training and costly annotating, for example with the aid of unsupervised learning or zero/few shots learning. While the previous state-of-the-art unsupervised model designed for news summarization (Zheng & Lapata, 2019) achieved only 4 ROUGE points improvement on out-of-domain data (i.e. scientific document summarization) over the lead baseline, I showed that we can **double the improvement** to 8 ROUGE points on out-of-domain data by **combining domain-specific discourse information with domain-agnostic graph-based algorithms**, achieving performance comparable to many state-of-the-art supervised approaches that are trained on hundreds of thousands of examples (Dong, Mircea, et al., 2021). With collaborators we also **constructed two scientific summarization datasets** that encourage research in domain adaptation from data-rich domains to data-scarce domains for single-document summarization (Meng, Thaker, Zhang, Dong, Yuan, Wang, & He, 2021) and multi-document summarization (Lu, Dong, et al., 2020) .

Research on domain generalization has led to a broad spectrum of methodologies, such as multi-task learning, transfer learning, domain adaptation, and meta-learning. With collaborators we have developed a general graph multi-task learning framework for neural sequence models, formulating the **multi-task problem as message passing over a graph neural network** (P. Liu, Dong*, Fu*, Qiu, & Cheung, 2019). By explicitly modeling the relationships between different tasks, our framework can enjoy performance gains from modeling mutually beneficial tasks in several multi-task and transfer learning settings.

3 Reducing factual hallucination in NLG

NLP consumers desire a reliable system that provides consistent and factual information, when interacting with AI agents such as conversation bots. Many companies have trained large-scale general neural language models (e.g. GPT3) that perform NLG tasks by highly abstracting the texts from billions of documents into its parameters to form the *parametric knowledge*. While delivering significant quality gains, these generation models that produce outputs word-by-word from scratch (seq2seq models), however, are prone to hallucinations with incorrect information (Maynez, Narayan, et al., 2020; Reiter, 2018).

To reduce hallucinations in the generation, with collaborators we propose simple solutions by constructing **pipeline of "generating then correcting" frameworks**, made up of conditional text generators and post-ad-hoc error correctors. This pipeline allows the system's output to **have improved summary factual correctness without sacrificing informativeness**. With collaborators we first propose an auto-regressive corrector based on distant-supervised learning (Cao, Dong, et al., 2020). To mitigate the hallucination inherent from auto-regressive seq2seq models, in (Dong, Wang, et al., 2020) we use a non-autoregressive neural model that is adapted from fact-checking QA model (Wang, Cho, et al., 2020) for factual correction on entities. In the most recent attempt, we propose to directly incorporate non-parametric knowledge, retrieved based on neural symbolic reasoning on knowledge graphs, to perform the factual correction. All the aforementioned neural-based correctors are lightweight and can be readily applied to any system-generated summaries without retraining the generation model.

This pipelined framework is effective in reducing factual errors despite they are notorious for cascading errors and increased computation. In many *monolingual* text-generation tasks where the source and target sequences have a considerable overlap such as text summarization or text simplification, a prominent alternative to neural-based seq2seq models is **text-editing models**. In Dong, Li, et al., 2019, we showed that generating the output **by predicting edit operations** applied to the source sequence, as opposed to generating outputs word-by-word from scratch (seq2seq models), resulted in higher performance with **better fact preserving of the inputs and better controllability of the outputs**.

4 Future directions

In future work, I plan to continue addressing the **fundamental challenges in NLP centered around robust generalization and faithful generation**. As an NLP researcher who is interested in building practical tools, I am primed to **collaborate with domain experts** outside computer science (e.g. researchers from social science and education) to **apply NLP techniques to real-life problems**, with a focus on practical and societal concerns such as reducing biases and promoting fairness.

Robust generalization To enable robust generalization to new domains, I plan to **develop new methods to separate domain agnostic vs. domain-specific features in NLG**. I am particularly interested in using **gradient-based methods for domain adaptation in natural language generation**. Gradient-based methods have been widely successful in various domain adaption settings for text classification (Sivaprasad, Goindani, et al., 2021). The intuition behind gradient-based methods is that the inverse gradients of domain-specific classifiers can force the neural networks to learn domain agnostic feature representations. I believe we can develop similar techniques for text generation tasks, for example, to better summarize e.g. scientific articles by transferring domain agnostic feature representations from data-rich domains such as news. I also plan to improve domain generalization in NLG tasks through **designing novel structures in multi-task learning (MTL)**. Multi-task learning has been wildly used in low-level NLP tasks mainly involved with text classification (P. Liu, Qiu, et al., 2016, 2017), how to modify these structures for sequential generation tasks with mutual beneficial information is largely under-explored. In addition, model parameters that perform well for one task are seldom the optimal solution for other tasks. It is worth investigating whether separating domain agnostic vs. domain-specific features can be achieved on the structure level, where the structural biases and typical heuristics that are specific to data-scare domains can be injected into the model structures for better generalization.

Faithful & factual generation Faithful generation is the fundamental requirement for conditional text generation models. I plan to continue my research to push the frontier in faithful text generation in the following dimensions. Before diving into techniques, one crucial dimension that is often being ignored is evaluation (Graham, 2015). Current evaluation metrics for faithful generation are broken (**Alexander2021summeval**; Kryscinski, Keskar, et al., 2019); efforts towards faithful generation are floating in the air before evaluation gets fixed. I believe the first step towards faithful generation is to **develop reliable and universal faithful generation measures** for different conditional generation tasks (e.g. GEM Mille, Dhole, et al., 2021). Towards faithful generation methodologies, I plan to investigate in **content planing and knowledge-grounded text generation**. Content planning has been wildly used in the pre-neural networks era (Cheung, Poon, et al., 2013; F. Liu, Flanigan, et al., 2015), where templates and knowledge graphs are learned as a medium to support reliable generation. How to integrate these knowledge-rich representations into neural models beyond the surface level (i.e. entities (Narayan, Zhao, et al., 2021)) remains a challenge. To advance research in this direction, I am particularly interested in integrating knowledge-grounded memories and syntactic templates into neural-based models for a controlled and faithful generation.

Biases reduction & fairness As an NLP researcher who is from an under-represented group, I plan to devote serious effort in applying NLP techniques to make a positive impact on **reducing biases and promoting fairness in social decision making such as hiring**. Language regularities biases in texts (e.g.racism and sexism) are often captured in word embeddings or parametric knowledge of neural-based models, which propagate into supervised downstream applications, such as information retrieval, text summarization, and web search. These automatic tools consequentially introduce biases and influence decision-making during resume screening for hiring. I am interested in revealing these biases and promoting fairness by fostering new collaborations with social scientists. We can use NLP to analyze the screening process to answer questions such as: How much do machine selections differ from human selections for candidate interviews? Are machine biases (e.g. man is to computer programmer and woman is to homemaker (Bolukbasi, Chang, et al., 2016)) distinguishable from unconscious human biases (e.g. whitened names are favored (Bertrand & Mullainathan, 2004))? In addition, word choice can trigger biases around education, maturity, and personality. Is the use of certain vocabularies in the applications more likely to result in certain individuals getting callbacks?

Education & mental health NLP also has a great potential to **improve learning efficiency in the classroom by providing feedback to both students and teachers**. Current NLP systems can provide formative feedback, ranging from lower-level advice (e.g. grammar and word choices) to higher-level suggestion (e.g. sentence structure and cohesion of discourse), to directly improve students' language skills. On the other side, NLP techniques can also be deployed for helping educators better identify and understand their students' cognition and mental health. The research in this direction is still nascent, leaving a large space for exploration. I am particularly interested in **collaborating with scientists in cognition and education to make a positive impact on learning with NLP techniques**. For example, by analyzing texts students produced NLP may help identify and predict students' mental states during learning, which can help teachers more effectively intervene and tailor instruction to individual needs. In addition, NLP tools can provide an automatic assessment of dynamic factors for students' learning (e.g. vocabulary knowledge and working memory); this information can help teachers to identify struggling students and provide appropriate assistance early on, resulting in an improved learning environment for students broadly.

References

Bertrand, M., & Mullainathan, S. (2004). Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American economic review*, *94*(4), 991–1013.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, *29*, 4349–4357.

Brandow, R., Mitze, K., & Rau, L. F. (1995). Automatic condensation of electronic publications by sentence selection. *Information Processing & Management*, *31*(5), 675–685.

Cao, M., Dong, Y., Wu, J., & Cheung, J. C. K. (2020). Factual error correction for abstractive summarization models. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Cheung, J. C. K., Poon, H., & Vanderwende, L. (2013). Probabilistic frame induction. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 837–846.

Dong, Y., Bhagavatula, C., Lu, X., Hwang, J. D., Bosselut, A., Cheung, J. C. K., & Choi, Y. (2021). On-the-fly attention modulation for neural generation. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 1261–1274. https://doi.org/10.18653/v1/2021.findings-acl.107

Dong*, Y., Grenander*, M., Cheung, J. C. K., & Louis, A. (2019). Countering the effects of lead bias in news summarization via multi-stage training and auxiliary losses. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 6019–6024. https://doi.org/10.18653/v1/D19-1620

Dong, Y., & Japkowicz, N. (2016). Threaded ensembles of supervised and unsupervised neural networks for stream learning. *Canadian conference on artificial intelligence*, 304–315.

Dong, Y., & Japkowicz, N. (2018). Threaded ensembles of autoencoders for stream learning. *Computational Intelligence*, *34*(1), 261–281.

Dong, Y., Li, Z., Rezagholizadeh, M., & Cheung, J. C. K. (2019). EditNTS: An neural programmer-interpreter model for sentence simplification through explicit editing. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3393–3402. https://doi.org/10.18653/v1/P19-1331

Dong, Y., Mircea, A., & Cheung, J. C. K. (2021). Discourse-aware unsupervised summarization for long scientific documents. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 1089–1102. https://www.aclweb.org/anthology/2021.eacl-main.93

Dong, Y., Shen, Y., Crawford, E., van Hoof, H., & Cheung, J. C. K. (2018). Banditsum: Extractive summarization as a contextual bandit. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3739–3748.

Dong, Y., Wang, S., Gan, Z., Cheng, Y., Cheung, J. C. K., & Liu, J. (2020). Multi-fact correction in abstractive text summarization. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 9320–9331. https://doi.org/10.18653/v1/2020.emnlp-main.749

Graham, Y. (2015). Re-evaluating automatic summarization with BLEU and 192 shades of ROUGE. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 128–137. https://doi.org/10.18653/v1/D15-1013

Kryscinski, W., Keskar, N. S., McCann, B., Xiong, C., & Socher, R. (2019). Neural text summarization: A critical evaluation. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 540–551. https://doi.org/10.18653/v1/D19-1051

Liu, F., Flanigan, J., Thomson, S., Sadeh, N., & Smith, N. A. (2015). Toward abstractive summarization using semantic representations. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1077–1086. https://doi.org/10.3115/v1/N15-1114

Liu, P., Dong*, Y., Fu*, J., Qiu, X., & Cheung, J. C. K. (2019). Learning multi-task communication with message passing for sequence learning. *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, 4360–4367. https://doi.org/10.1609/aaai.v33i01.33014360

Liu, P., Qiu, X., & Huang, X. (2016). Recurrent neural network for text classification with multi-task learning. *IJCAI*.

Liu, P., Qiu, X., & Huang, X. (2017). Adversarial multi-task learning for text classification. *ACL (1)*.

Lu, Y., Dong, Y., & Charlin, L. (2020). Multi-XScience: A large-scale dataset for extreme multi-document summarization of scientific articles. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 8068–8074. https://doi.org/10.18653/v1/2020.emnlp-main.648

Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Meng, R., Thaker, K., Zhang, L., Dong, Y., Yuan, X., Wang, T., & He, D. (2021). Bringing structure into summaries: A faceted summarization dataset for long scientific documents. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 1080–1089. https://doi.org/10.18653/v1/2021.acl-short.137

Mille, S., Dhole, K., Mahamood, S., Perez-Beltrachini, L., Gangal, V., Kale, M., van Miltenburg, E., & Gehrmann, S. (2021). Automatic construction of evaluation suites for natural language generation datasets. *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*. https://openreview.net/forum?id=CSi1eu_2q96

Narayan, S., Zhao, Y., Maynez, J., Simoes, G., Nikolaev, V., & McDonald, R. (2021). Planning with learned entity prompts for abstractive summarization. *arXiv preprint arXiv:2104.07606*.

Nenkova, A. (2005). Automatic text summarization of newswire: Lessons learned from the document understanding conference. *AAAI*, *5*, 1436–1441.

Reiter, E. (2018). A structured review of the validity of BLEU. *Computational Linguistics*, *44*(3), 393–401. https://doi.org/10.1162/coli_a_00322

Sivaprasad, S., Goindani, A., Garg, V., & Gandhi, V. (2021). Reappraising domain generalization in neural networks. *arXiv preprint arXiv:2110.07981*.

Wang, A., Cho, K., & Lewis, M. (2020). Asking and answering questions to evaluate the factual consistency of summaries. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Reinforcement learning* (pp. 5–32). Springer.

Zheng, H., & Lapata, M. (2019). Sentence centrality revisited for unsupervised summarization. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 6236–6247.