From Perceptron to Cognition: Towards Next Generation Intelligent Agent

Xiaoyu Shen

My research goal is centered around advancing the frontier of machine learning (ML) and natural language processing (NLP) to build next generation *personalized intelligent agents*. In particular, this requires developing agents that can interact with humans through natural languages to figure out the needs, conduct various tasks and provide explanations.

Background Over the last decades, ML research has been featured with developing isolated systems to solve task-specific problems such as object detection, sentiment analysis and machine translation. With the wide adoption of deep neural networks and growing scale of annotated datasets, many systems are able to achieve close-to-human performances in their targeted domains. Nonetheless, they are mostly trained via supervised learning to predict human-provided labels without knowing how the labels are derived in human thoughts. In this sense, the systems themselves are running only at the "perceptron" level: they can perceive input signals, predict the output labels by learning feature mappings from the training data, but lack the mental process to understand the physical world that the task is grounded on. As a result, they are limited as narrow experts in their strictly defined input-output space, sensitive to domain shifts and hard to provide explanations to their output.

Recently, the outburst of NLP applications powered by large language models (LLMs) have demonstrated great potentials towards understanding the world at the "cognition" level [33, 32]. Languages, being the cornerstone of human intelligence, are a foundational tool to formulate and communicate an extraordinarily broad range of thoughts including our abstract and general world knowledge, our common believes and our approaches to reason about specific problems [6, 20]. There has also been a strong belief that the structure of a language influences its speakers' worldview or cognition, and thus individuals' languages determine or shape their perceptions of the world of human intelligence [12]. Therefore, LLMs trained on Internet-level language text are often considered to be equipped with basic cognitive capabilities. They are able to follow human instructions, interact with various plug-ins and provide justifications to their output, all of which through the form of natural languages [39].

Challenges Despite the promising achievements, existing applications are still far from being reliable intelligent agents. Three noticeable challenges are (1) *dominance of English language*: the overwhelmingly majority of Internet data is written in English, the resulting trained LLMs are thus much better at English than other languages. This makes it hard for other language users to obtain the same quality of service, which can aggregate the regional development imbalance; (2) *lack of domain-specific knowledge*: Existing LLMs are good at commonsense knowledge but struggles at specialized domains where highly professional knowledge is required. A lot of domain-specific knowledge is stored in private databases and updated frequently, which makes it hard for the knowledge to be fully absorbed by a static LLM; and (3) *difficulty at providing interpretable and verifiable outputs*: Though LLMs are able to provide explanations for their output, these explanations are often inconsistent with the actual generation process. It is also very hard to verify the correctness as the generated references from LLMs are often inaccurate [10].

In the following section, I will first present my prior research experience and highlight how they could help achieve my research goal (sections 1 to 4), then elaborate on the future directions to address the above-mentioned three challenges (§5).

1 Latent-Variable Model

During my PhD, I devoted a lot of efforts in developing efficient optimization algorithms for latent-variable models in text generation [11, 15]. Humans do not produce natural languages out of blue. The mental process of language generation involves complicated interactions of multifarious factors, obtaining annotations of all factors is extremely difficult, if not impossible. Latent-variable models provide a powerful tool to build personalized intelligent agents by mimicking this mental process: They specify a prior distribution from which latent factors are drawn, and a likelihood distribution from which the actual text is sampled conditioned on the latent factors. The factors are latent and automatically inferred during the model training process, so that no additional manual annotation is required.

Improving Diversity and Consistency As the injection of latent variables brings additional stochasticity, it can be naturally used to improve the diversity of model output. In [23], we integrate Gaussian distributed latent variables into conversational dialogue generation, and infer them using variational autoencoders. The resulting model can significantly improve the diversity of generated responses without sacrificing other features. As the integration of latent variable makes the training process unstable, we propose an optimization algorithm to stabilize the training, in which we theoretically prove the superior upper bound of the proposed optimization algorithm [21, 24]. In a follow-up work [22], we further show that latent variables can be used to enhance the consistency with previous and next utterances in a dialogue flow, where we optmize an upper bound of mutual information between dialogue utterances and continuous latent variables.

Improving Interpretability and Controllability Another advantage of latent-variable models is that they allow easier interpretation to model output. We can also control the generation by specifying values to these latent variables beforehand. In [28], we use latent variables to capture the word alignment so that every generated word has explicit source attribution. In [25], we propose a framework to enable controllable content selection in text generation problems. The trade-off between controllability and generation quality can be adjusted through an approximation of conditional mutual information. In [19], we define the segmentation and correspondence of model generations as latent variables, and propose an efficient optimization algorithm by simplifying the generation process as semi-markov models. The resulting model achieves close-to-zero error rates in two popular benchmarks and provides fully interepretable output structures.

2 Low-Resource Training

The second thread of my work focuses on low-resource training [40, 26, 27], which aims to solve NLP problems with limited amounts of human annotations. In reality, high-quality human annotations are usually very costly to obtain. For certain domains, they are even infeasible due to legal issues and lack of data or domain experts. This challenge is further exacerbated in training personalized intelligent agents as they are expected to be involved in an open-ended set of tasks tailored for various types of users. It is thereby very important to make the most of the limited human annotations.

Data Augmentation There are two general ways of data augmentation in NLP tasks: (1) Given a labeled input utterance, augment the data by generating more utterances with the same label; and (2) Given an unlabeled input, augment the data by generating a pseudo label for it. For the first, we conduct extensive empirical experiments to compare both word-level and sentence-level transformations. We find that the effectiveness of different data augmentation schemes depends on the nature of the dataset under consideration [36, 37, 5]. Based on it, we develop EasyAug, a data augmentation platform that provides popular choices of augmentation approaches for practitioners to easity compare [13]. For the second, we propose MSR, a meta-learning-based training framework that adaptively transforms weak pseudo labels into stronger, informative labels. It has achieved state-of-the-art performances across 8 diverse set of NLP tasks [41]. In a follow-up work, we further investigated existing weakly supervised learning methods and show their claimed advantages are significantly overestimated, which has won the *best theme paper award in ACL 2023* [42].

Active Learning Data augmentation augments data based on existing annotations, while active learning actively selects most informative samples to be annotated that can lead to biggest performance improvement. In [2], we present DART: a lightweight quality-suggestive annotation tool that selects samples based on the bidirectional reconstruction loss, which we show can significantly reduce the annotation cost in order to achieve the same performance. The tool has won the *best demo paper award at COLING 2020*. In [4], we further propose a clustering based method to suggest initial data points for annotation before the active learning stage, which can reduce the number of required annotations by half. We also organized a challenge to search for optimal active learning strategies in natural language generation [3].

3 Commercial Intelligent Agents

The third thread of my work lies in developing commercially usable intelligent agent systems. Different from laboratory environments, real user traffic is of a much wider scope and contains noise from different levels. The problem setup and development process are also dynamic, require active refinement, collaborations from multiple parties and a more comprehensive view of the whole system. Productionizing experience in the industry is crucial for future work if the goal is to build usable intelligent agents in real user scenarios.

Chitchat Conversational Agent for Wechat Starting from late 2018, I started collaborating with the wechat AI team on developing chitchat conversational agents. To address the frequently occurred coreference and information omission in multi-turn daily chats, we proposed a novel utterance rewriting technique to converse multi-turn dialogue flows into single-turn queries, and demonstrated improvements in online A/B tests [30]. We also proposed the first technique that allows utilizing non-conversational text to diversity dialogues [31]. Built upon these, we developed the state-of-the-art Chinese chatbot in the movie domain which could converse with humans for over 10 turns [29]. To prevent online user adversaries, we further pre-trained a robust language model specific, which can help significantly reduce performance drop under noisy user inputs [32].

Product Question Answering for Alexa Shopping In September 2020, I joined Amazon Alexa AI to work on product question answering. Unlike chitchat conversational agents, product question answering cares more about accuracy of information seeking than entertainment. We started from developing a simple prototype system involving only semi-structured information on simulated user questions [18], then expanded to heterogeneous information on realistic user questions [17]. Starting from the second year, the system has contributed to the most query success rates and replaced the old ensemble-based systems. We pre-trained our system on billions of internal question-answer pairs with weak supervision [27], and generalized it to 12 non-English languages [16]. The system has been the main answer provider in Amazon serving hundreds of millions of users.

4 Applications in Other Domains

Finally, I also spent time developing applications in other domains. For example, in [1], we applied graphical inference with loopy propagation to study large-scale erosion of genomic privacy over time. I was the main contributor to implement the whole system and run all experiments. As the main student supervisor, in [34, 35], we proposed a novel structured attention to integrate abstract-syntax tree in source code summarization. To enable efficient computation, we further implemented a novel gather with decomposed coordinate format to reduce the computational complexity from quadratic to linear. In the legal domain, We maintained state-of-the-art fact-article mathing and legal judgement prediction model in the legal domain [8, 9, 7]. We also continued to pre-train a large language model on tens of millions of question-answer pairs to support real-time consultation in Chinese laws [1].

Getting exposed to and being experienced with various domains help establish a broader view in model development. This is important in developing personalized intelligent agents as it requires knowledge from multiple domains. These experiences will also help me efficiently talk to domain experts, understand a domain and propose domain-specific solutions in the future.

[1] `https://github.com/davidpig/lychee_law`

5 Future Directions

My prior research has laid out a good foundation to build the next generation personalized intelligent agent that can think, reason and talk in a human-like way. The recent rapid development of LLMs have shown that the model performance can be steadily improved as the growing model and data size. Moving forward, my mission to explore to which extend LLMs can lead us towards the goal and how my expertise could help reduce the gap. With this in mind, I plan on pushing forward in the following research areas:

Cross-lingual generalization The imbalanced distribution of language resources online makes it very inefficient to re-train intelligent agents from scratch for every new language [38]. Even for the same language, there exists different locales and dialects which evolve over time, which makes cross-lingual generalization a crucial task. Previously I have supervised a student project that shows the model weights trained on English carry useful information on other languages as well, even when we directly plug them it with vocabularies from new languages [14]. In a collaborative project with Masakhane nlp, we have similar observations on African languages that have never been seen during the pre-training stage of existing language models. These indicate great potentials in speeding up the generalization to new languages with existing pre-trained English-centric models. Concrete further steps are (1) repeating our current experiments on LLMs and identify the minimal required corpus to enable successful transfer. This is related to my previous research on low-resource training (data selection in active learning especially); (2) exploring word and grammar alignment in LLM. They have been frequently done in smaller LMs but never in LLMs yet. I am currently leading a project with a PhD student from Saarland university in doing so.

From general to domain expert Existing pre-trained LLMs are general intelligent agents that are remarkably good at tasks requiring basic knowledge, but are far from good domain experts. In a recent legal assistant project in Nanjing University, we find that ChatGPT often hallucinates wrong law articles, news events, and make wrong judgement predictions, even though the answers look plausible themselves. They also tend to provide general, shallow suggestions that are hardly usable in real-life legal consultancy. We have been working to collect high-quality legal-domain corpus to train a domain expert in Chinese laws. The currently released model only supports question-answering tasks, but we are working to support more diverse interactive instructions. I am also leading a project exploring whether we can use structurally organized textbooks to more efficiently adapt LLMs from general do domain experts. Intuitively humans learn new domain knowledge with well-designed textbooks from basic to advanced concepts. LLMs might also be able to mimic this learning process of humans. If LLMs can learn in a similar way as humans, it would become much easier since small structured textbooks are easier to obtain and more efficient to utilize than huge unstructured domain-specific corpus. Ideally we should come up with a systematic way to sequentially adapt LLMs to (1) memorize domain-specific concepts; (2) understand these concepts; and (3) apply them to concrete tasks.

Provide interpretable and verifiable output To develop usable and trustful intelligent agents, it is necessary that their output is interpretable and verifiable. There are generally two paths to achieve this. The first path adopts pri-explain, where the model explicitly follows a series of steps to come to the final answer instead of directly decodes word by word. The intermediate steps can be thought process, tool calling, web search, etc, which can be used to interpret and verify the model output. This is very relevant with my research experience in latent-variable models, where these intermediate steps are inferred as latent variables. The risk of this path is that it is very hard to make all decision steps explicit and the inference will be significantly slowed down. As the human mental process naturally contains complex procedures hard to be expressed by natural languages, we might need to design more expressive symbols to represent the decision steps. The second path adopts post-explanation, where the model generates explanations for its output after the output has been generated. Its advantage would be that the original generating process will not be affected so that the inference latency will stay the same. However, the accuracy of post-explanation is also lower. As the output generation and explanation generation are two separate steps, there is no guarantee that the generated explanation coincides with the output generation process. The future research would very likely need to combine both paths as a trade-off. It would also be super interesting to connect it with neural science as there have been studies showing that human brains think in a similar way: the “conscious” thinking is doing pri-explanation and the “sub-conscious” thinking is doing post-explanation. Developing AI might also help us understand humans better.

In conclusion, these future directions are both challenging and exciting, and I believe that the work along these directions will take us a significant step closer to the next generation personalized intelligent agents that can be reliably used in our daily lives.

References

[1] Michael Backes, Pascal Berrang, Mathias Humbert, Xiaoyu Shen, and Verena Wolf. Simulating the large-scale erosion of genomic privacy over time. *IEEE/ACM transactions on computational biology and bioinformatics*, 15(5):1405–1412, 2018.

[2] Ernie Chang, Jeriah Caplinger, Alex Marin, Xiaoyu Shen, and Vera Demberg. Dart: A lightweight quality-suggestive data-to-text annotation tool. In *Proceedings of the 28th International Conference on Computational Linguistics: System Demonstrations*, pages 12–17, 2020.

[3] Ernie Chang, Xiaoyu Shen, Alex Marin, and Vera Demberg. The selectgen challenge: Finding the best training samples for few-shot neural text generation. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 325–330, 2021.

[4] Ernie Chang, Xiaoyu Shen, Hui-Syuan Yeh, and Vera Demberg. On training instance selection for few-shot neural text generation. In *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)*, pages 8–13, 2021.

[5] Ernie Chang, Xiaoyu Shen, Dawei Zhu, Vera Demberg, and Hui Su. Neural data-to-text generation with lm-based text augmentation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 758–768, 2021.

[6] Nick Chater and Christopher D Manning. Probabilistic models of language processing and acquisition. *Trends in cognitive sciences*, 10(7):335–344, 2006.

[7] Junyun Cui, Xiaoyu Shen, Feiping Nie, Zheng Wang, Jinglong Wang, and Yulong Chen. A survey on legal judgment prediction: Datasets, metrics, models and challenges. *arXiv preprint arXiv:2204.04859*, 2022.

[8] Jidong Ge, Yunyun Huang, Xiaoyu Shen, Chuanyi Li, and Wei Hu. Learning fine-grained fact-article correspondence in legal cases. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3694–3706, 2021.

[9] Yunyun Huang, Xiaoyu Shen, Chuanyi Li, Jidong Ge, and Bin Luo. Dependency learning for legal judgment prediction with a unified text-to-text transformer. *arXiv preprint arXiv:2112.06370*, 2021.

[10] Nelson F Liu, Tianyi Zhang, and Percy Liang. Evaluating verifiability in generative search engines. *arXiv preprint arXiv:2304.09848*, 2023.

[11] Aditya Mogadala, Xiaoyu Shen, and Dietrich Klakow. Integrating image captioning with rule-based entity masking. *arXiv preprint arXiv:2007.11690*, 2020.

[12] Harriet Joseph Ottenheimer and Judith MS Pine. *The anthropology of language: An introduction to linguistic anthropology*. Cengage Learning, 2018.

[13] Siyuan Qiu, Binxia Xu, Jie Zhang, Yafang Wang, Xiaoyu Shen, Gerard De Melo, Chong Long, and Xiaolong Li. Easyaug: An automatic textual data augmentation platform for classification tasks. In *Companion Proceedings of the Web Conference 2020*, pages 249–252, 2020.

[14] Lei Shen, Shuai Yu, and Xiaoyu Shen. Is translation helpful? an empirical analysis of cross-lingual transfer in low-resource dialog generation. *arXiv preprint arXiv:2305.12480*, 2023.

[15] Xiaoyu Shen. Deep latent-variable models for text generation. *arXiv preprint arXiv:2203.02055*, 2022.

[16] Xiaoyu Shen, Akari Asai, Bill Byrne, and Adria De Gispert. xPQA: Cross-lingual product question answering in 12 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 103–115, Toronto, Canada, July 2023. Association for Computational Linguistics.

[17] Xiaoyu Shen, Gianni Barlacchi, Marco Del Tredici, Weiwei Cheng, Bill Byrne, and Adrià de Gispert. Product answer generation from heterogeneous sources: A new benchmark and best practices. In *Proceedings of the Fifth Workshop on e-Commerce and NLP (ECNLP 5)*, pages 99–110, 2022.

[18] Xiaoyu Shen, Gianni Barlacchi, Marco Del Tredici, Weiwei Cheng, and Adrià de Gispert. semipqa: A study on product question answering over semi-structured data. In *Proceedings of the Fifth Workshop on e-Commerce and NLP (ECNLP 5)*, pages 111–120, 2022.

[19] Xiaoyu Shen, Ernie Chang, Hui Su, Cheng Niu, and Dietrich Klakow. Neural data-to-text generation via jointly learning the segmentation and correspondence. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7155–7165, 2020.

[20] Xiaoyu Shen, Youssef Oualil, Clayton Greenberg, Mittul Singh, and Dietrich Klakow. Estimation of gap between current language models and human performance. *Proc. Interspeech 2017*, pages 553–557, 2017.

[21] Xiaoyu Shen and Hui Su. Towards better variational encoder-decoders in seq2seq tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[22] Xiaoyu Shen, Hui Su, Wenjie Li, and Dietrich Klakow. Nexus network: Connecting the preceding and the following in dialogue generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4316–4327, 2018.

[23] Xiaoyu Shen, Hui Su, Yanran Li, Wenjie Li, Shuzi Niu, Yang Zhao, Akiko Aizawa, and Guoping Long. A conditional variational framework for dialog generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 504–509, 2017.

[24] Xiaoyu Shen, Hui Su, Shuzi Niu, and Vera Demberg. Improving variational encoder-decoders in dialogue generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[25] Xiaoyu Shen, Jun Suzuki, Kentaro Inui, Hui Su, Dietrich Klakow, and Satoshi Sekine. Select and attend: Towards controllable content selection in text generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 579–590, 2019.

[26] Xiaoyu Shen, Svitlana Vakulenko, Marco Del Tredici, Gianni Barlacchi, Bill Byrne, and Adrià de Gispert. Low-resource dense retrieval for open-domain question answering: A comprehensive survey. *arXiv preprint arXiv:2208.03197*, 2022.

[27] Xiaoyu Shen, Svitlana Vakulenko, Marco Del Tredici, Gianni Barlacchi, Bill Byrne, and Adrià de Gispert. Neural ranking with weak supervision for open-domain question answering: A survey. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1691–1705, 2023.

[28] Xiaoyu Shen, Yang Zhao, Hui Su, and Dietrich Klakow. Improving latent alignment in text summarization by generalizing the pointer generator. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3762–3773, 2019.

[29] Hui Su, Xiaoyu Shen, Zhou Xiao, Zheng Zhang, Ernie Chang, Cheng Zhang, Cheng Niu, and Jie Zhou. Moviechats: Chat like humans in a closed domain. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 6605–6619, 2020.

[30] Hui Su*, Xiaoyu Shen*, Rongzhi Zhang, Fei Sun, Pengwei Hu, Cheng Niu, and Jie Zhou. Improving multi-turn dialogue modelling with utterance rewriter. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 22–31, 2019.

[31] Hui Su*, Xiaoyu Shen*, Sanqiang Zhao, Zhou Xiao, Pengwei Hu, Randy Zhong, Cheng Niu, and Jie Zhou. Diversifying dialogue generation with non-conversational text. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7087–7097, 2020.

[32] Hui Su*, Weiwei Shi*, Xiaoyu Shen*, Zhou Xiao, Tuo Ji, Jiarui Fang, and Jie Zhou. Rocbert: Robust chinese bert with multimodal contrastive pretraining. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 921–931, 2022.

[33] Hui Su, Xiao Zhou, Houjing Yu, Yuwen Chen, Zilin Zhu, Yang Yu, and Jie Zhou. Welm: A well-read pre-trained language model for chinese. *arXiv preprint arXiv:2209.10372*, 2022.

[34] Ze Tang, Chuanyi Li, Jidong Ge, Xiaoyu Shen, Zheling Zhu, and Bin Luo. Ast-transformer: Encoding abstract syntax trees efficiently for code summarization. In *2021 36th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pages 1193–1195. IEEE, 2021.

[35] Ze Tang, Xiaoyu Shen, Chuanyi Li, Jidong Ge, Liguo Huang, Zhelin Zhu, and Bin Luo. Ast-trans: Code summarization with efficient tree-structured attention. In *Proceedings of the 44th International Conference on Software Engineering*, pages 150–162, 2022.

[36] Liqiang Wang, Xiaoyu Shen, Gerard de Melo, and Gerhard Weikum. Cross-domain learning for classifying propaganda in online contents. *arXiv preprint arXiv:2011.06844*, 2020.

[37] Binxia Xu, Siyuan Qiu, Jie Zhang, Yafang Wang, Xiaoyu Shen, and Gerard de Melo. Data augmentation for multiclass utterance classification–a systematic study. In *Proceedings of the 28th international conference on computational linguistics*, pages 5494–5506, 2020.

[38] Qingyu Zhang, Xiaoyu Shen, Ernie Chang, Jidong Ge, and Pengke Chen. Mdia: A benchmark for multilingual dialogue generation in 46 languages. *arXiv preprint arXiv:2208.13078*, 2022.

[39] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

[40] Yang Zhao, Xiaoyu Shen, Wei Bi, and Akiko Aizawa. Unsupervised rewriter for multi-sentence compression. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2235–2240, 2019.

[41] Dawei Zhu, Xiaoyu Shen, Michael Hedderich, and Dietrich Klakow. Meta self-refinement for robust learning with weak supervision. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1043–1058, 2023.

[42] Dawei Zhu, Xiaoyu Shen, Marius Mosbach, Andreas Stephan, and Dietrich Klakow. Weaker than you think: A critical look at weakly supervised learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14229–14253, Toronto, Canada, July 2023. Association for Computational Linguistics.