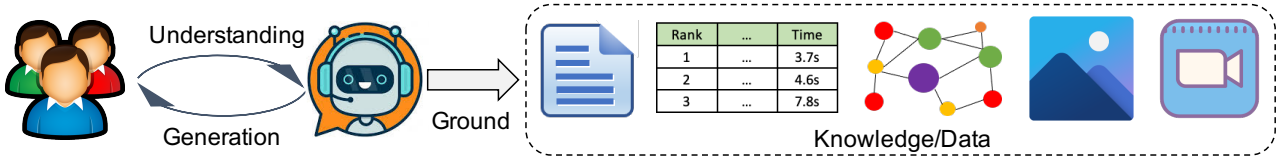
RESEARCH STATEMENT

Wenhu Chen (wenhuchen@cs.ucsb.edu)

A long-standing goal of artificial intelligence is to build machines like Apple Siri, Amazon Alexa that can ground on world knowledge to interact with a human to accomplish desired tasks. As the world-knowledge is distributed in diverse forms of data including text, graphs, tables, charts, images, and videos, a core problem to build a system that can understand and interact with these diversified data. My research goal is to design **general** and **efficient** machine learning models that can generalize to different settings with low sample and computation complexity. My research scope can be roughly divided into the following two categories:

- Natural language Understanding: how to understand human language input and ground it on real-world data to perform tasks like information retrieval, question answering, fact-checking, etc.
- Natural Language Generation: how to generate coherent and faithful natural language from real-world data to communicate with human.

In the future, I look forward to advancing natural language models to deal with these two fundamental problems. I also enjoy interdisciplinary research and collaboration with experts from different fields such as data mining, programming language, security, data science, social computing, etc.



1 Natural Language Understanding

My research in natural language understanding is focused on designing question answering and fact-checking models that can deal with diversified real-world data.

Extractive Question Answering Extractive question answering considers the problem of answering a given question by searching text span from the given data. The existing QA datasets are restricted to dealing with text data like passages. However, the real-world knowledge is distributed over heterogeneous forms, it's not ideal to consider the text as the only corpus. To simulate a more realistic setting, I proposed a dataset HybridQA [1] to answer a given question based on a set of tabular and textual data, where the model needs to aggregate information from two forms to find the answer. To handle such a challenge, I designed a model that can iteratively hop between two forms to accumulate evidence and finally predict an answer span. Following this line, I proposed an open-domain HybridQA [2], where the supporting tabular and textual data are not given, and the model needs to retrieve them from the web to answer a given question. The new problem is challenging due to the cross-form dependency during the retrieval process. The standard approach uses iterative retrieval, which is expensive due to multiple rounds of encoding and searching. To address these issues, an 'early' fusion mechanism was devised to group highly related tabular and textual units offline as a semantic cluster and jointly encode them. Instead of retrieving an individual unit, the model can directly retrieve these clusters as a whole. The new approach not only improves the retrieval accuracy but also decreases the computation cost greatly.

Natural Language Inference Natural language inference considers the problem of predicting whether a textual hypothesis is entailed by a premise, which has been widely adopted in applications like fact verification, fake news detection, etc. However, the existing paradigm has two limitations: 1) the current datasets only consider the textual data as the premise, 2) the inference is limited to the semantic level, without involving symbolic/logical operations. To broaden the scope, I proposed table-based natural language inference [3]. The new problem considers semi-structured web table data as the premise and requires symbolic reasoning (counting,

summation, etc) to predict the verdict of a hypothesis. We experimented with semantic parsing-based and NLI-based models to separately solve the problem and both models perform equally well on the dataset. We further used a graph neural network to combine the merits in symbolic reasoning and semantic reasoning. This project received over 200+ stars on GitHub within a year and inspired many follow-up studies by institutes including Microsoft, Google, IBM, etc. Besides, I also proposed the video-text natural language inference [4], where the given premise becomes a video captioned with subtitles. The new dataset involves both visual and textual modalities, which poses great challenges to existing models. We used different models that can align video frames with text snippets to perform joint reasoning. The proposed two new tasks greatly enriched the existing natural language inference family to investigate more complex open-world scenarios.

Visual Question Answering visual question answering aims to answer a question from a given image. A popular approach to solve visual question answering is module network due to its strong explainability and compositionality. This method first defines a set of functions and then associate hand-crafted neural networks with the pre-defined functions. However, this method suffers from scalability and generalization due to its hand-crafting procedure. I proposed the meta-module network [5] to learn the neural modules with meta-learning algorithm from the data to avoid hand-crafting. The proposed model demonstrates a strong generalization ability to accommodate a larger set of functions and improves the final accuracy.

2 Natural Language Generation

My research in natural language generation is focused on designing generation models that are coherent and consistent with real-world knowledge.

Data-to-Text Generation The problem of data-to-text generation considers producing the human language from given data in various forms like graphs, tables, charts, etc. It has been widely used in different applications to produce documents, advertisements, reports, explanations, help messages, etc. The previous studies proposed ad-hoc models for a specific domain or data form. These models need to be trained from scratch with domain-specific labeled data and cannot generalize well to the other domains. To design a universal data-to-text model that can generalize to a wide range of tasks, we first devised a unified graph-to-text model KGPT [6]. Similar to GPT-2, KGPT model is pre-trained on large-scale automatically constructed graph-to-text corpus from the web. After seeing millions of examples from the web, the model acquires a strong generalization ability to adapt to various data-to-text tasks with as few as 100 examples to achieve similar performance as the previous models seeing tens of thousands of examples.

Another direction I explored is to enhance the reasoning capability of the existing data-to-text generation models so that the model can derive more implicit facts from the given data. For example, the data contains 'The U.S obtain 3 gold medals, 2 silver medals, 1 bronze medal', the model needs to generate more high-level summarization like 'The U.S obtained a total of 6 medals' with logical operations. Such reasoning capability is non-existing in the current generation models, therefore, we designed a coarse-to-fine generation framework to first generate a coarse-grained template with holes over the logic-related words and then fill in these holes with a fine-grained generator. The proposed method can improve the logical correctness over existing models.

Dialog Generation The problem of task-oriented dialog considers the problem of conversating with humans to accomplish certain tasks like hotel booking, customer service, etc. The existing system adopts a pipeline framework to understand human input, and then predict a structured dialog action and then convert the predicted dialog action to natural language response. The previous approaches aim to represent dialog act into vectors and feed it as auxiliary inputs to influence the response generation, which only provides weak controllability to ensure the accuracy of generated response. To improve the controllability of the generation process, I designed a disentangled transformer model [7]. Unlike the standard transformer, the disentangled transformer can disentangle its different attention heads to represent a designated semantics in the dialog action space during training. For example, a specific attention head can be in charge of 'hotel' or 'restaurant' to control the response to fall into its semantics. In the test time, given a dialog action, the model toggle the attention heads to reflect

the semantics of given dialog actions to generate more accurate responses. The proposed model is shown to greatly improve both automatic and human evaluation score by a large margin.

Visual Captioning The problem of visual captioning aims to generate natural language captions grounded on visual inputs. In video captioning, the generation model needs to ground on a sequence of actions in the video to generate fine-grained descriptions. To better such sequential actions, we proposed hierarchical reinforcement learning [8] to break down the action sequences into a set of individual actions and use them as the sub-goals. The model will assign rewards in the intermediate sub-goal steps to resolve the reward sparsity problem.

To investigate whether the current captioning evaluation metrics like BLEU, CIDEr are good metrics for visual-text grounding, I conducted human studies and found that these metrics have low consensus with human perception. Through the experiments, I showed that using these metrics as rewards in reinforcement learning can make the model collapse to generate garbage output in favor of a certain metric. Therefore, I designed an inverse reinforcement learning [9] to derive a model-based metric from the data through a learning process. The learned reward function is shown more robust to help the model better explore the output space to generate more human-like captions without biases to favor certain metrics.

3 Interdisciplinary Research

Besides the above-mentioned problems, I'm also interested in collaborating in other areas like machine learning, data mining, time series forecasting, and security. For example, I designed a variational inference framework [10] to automatically learn the importance of different vocabulary words in different natural language processing tasks. By leveraging the word-level importance, we can filter out words with low importance weight and compress different NLP models significantly. We found that on the most popular text classification tasks, the vocabulary can be compressed by over 90% to maintain the same performance. For data mining, I also collaborated with other researchers to apply the advanced BERT model to mine algorithms and find their relations from scientific literature [11]. By combining natural language processing techniques, the mining accuracy can be greatly improved. For time series forecasting, I collaborated with a junior PhD student to apply the transformer model to predict trends in time series data [12]. In order to handle the long sequence of history, we devised sparse attention in the transformer to greatly reduce the attention computation complexity. We demonstrate that the deep forecasting model can achieve significant improvement over the statistical models on time series forecasting. For security, I'm interested in designing algorithms to generate adversarial examples to break the current deep learning systems and using these adversarial examples to improve the model's robustness before model deployment.

4 Future Work

In the future, I would continue to develop general and powerful models that can ground on diverse forms of open-world data and handle real-world complex scenarios in real-world applications. My life-long goal is to make impactful research that can be applied to the industries to create real values for society. To achieve this goal, I would lay out my research plan in these three directions: 1) designing universal models: like KPGT model, I plan to design general-purpose models that can be applied across different domains and different settings with strong generalization capability. Such universal models can greatly lower the cost of preservation when deployed in the industries. 2) few-shot/zero-shot learning: one of the most severe barriers for applying deep models into industries is the data hunger problem, I plan to design a more sample-efficient algorithm to understand a task with only a handful of examples to achieve good performance. I recently work on unsupervised question answering models which can synthesize good QA pairs for training [13]. By lowering the annotation costs, the deep model becomes more applicable in the industries. 3) social responsibility: once deploying deep models into industries, we might face serious issues in ethics, fairness, justice, etc. I'm eager to solve these problems to make the model more robust and contribute to our society more responsibly.

Meanwhile, I enjoy collaborating with scientists and domain experts of different backgrounds for interdisciplinary research in various domains like information retrieval, computational social science, and programming language, etc. I'm excited to apply my knowledge and expertise to solve problems in other areas. For example, I'm interested in collaborating with IR researchers to combine the neural-based retrieval system with BM25

retrieval systems to achieve better performances. I'm excited about collaborating with computational social science researchers to improve the current fake news and misinformation detection with the help of the most recent NLP techniques. I'm also delighted to work with programming language researchers to design a system that can find convert human language into logical forms to execute functions or access databases, etc.

References

[1] Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Wang. Hybridqa: A dataset of multi-hop question answering over tabular and textual data. *Proceedings of EMNLP (Findings)*, 2020.

[2] Wenhu Chen, Ming-Wei Chang, Schlinger Eva, and William W. Cohen. Open question answering over tables and text. *Arxiv*, 2020.

[3] Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. Tabfact: A large-scale dataset for table-based fact verification. In *International Conference on Learning Representations*, 2020.

[4] Jingzhou Liu, Wenhu Chen, Yu Cheng, Zhe Gan, Licheng Yu, Yiming Yang, and Jingjing Liu. Violin: A large-scale dataset for video-and-language inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10900–10910, 2020.

[5] Wenhu Chen, Zhe Gan, Linjie Li, Yu Cheng, William Wang, and Jingjing Liu. Meta module network for compositional visual reasoning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, March 2021.

[6] Wenhu Chen, Yu Su, Xifeng Yan, and William Yang Wang. Kgpt: Knowledge-grounded pre-training for data-to-text generation. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020.

[7] Wenhu Chen, Jianshu Chen, Pengda Qin, Xifeng Yan, and William Yang Wang. Semantically conditioned dialog response generation via hierarchical disentangled self-attention. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3696–3709, 2019.

[8] Xin Wang, Wenhu Chen, Jiawei Wu, Yuan-Fang Wang, and William Yang Wang. Video captioning via hierarchical reinforcement learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[9] Xin Wang*, Wenhu Chen*, Yuan-Fang Wang, and William Yang Wang. No metrics are perfect: Adversarial reward learning for visual storytelling. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 899–909, 2018.

[10] Wenhu Chen, Wenhan Xiong, Xifeng Yan, and William Yang Wang. Variational knowledge graph reasoning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1823–1832, 2018.

[11] Hanwen Zha, Wenhu Chen, Keqian Li, and Xifeng Yan. Mining algorithm roadmap in scientific publications. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1083–1092, 2019.

[12] Shiyang Li, Xiaoyong Jin, Yao Xuan, Xiyou Zhou, Wenhu Chen, Yu-Xiang Wang, and Xifeng Yan. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. In *Advances in Neural Information Processing Systems*, pages 5243–5253, 2019.

[13] Liangming Pan, Wenhu Chen, Wenhan Xiong, Min-Yen Kan, and William Yang Wang. Unsupervised multi-hop question answering by question generation. *arXiv preprint arXiv:2010.12623*, 2020.