Research Statement

Junjie Hu

My research agenda is centered around advancing the frontier of artificial intelligence (AI) to support **multilingual human-machine communication**. In particular, my research involves the development of **natural language processing (NLP)** and **machine learning (ML)** techniques to handle **diverse signals** from the dynamic environment, as shown in Figure 1. This human-machine communication process requires NLP systems to understand complex data of a variety that might not be available for system training. However, current NLP systems powered by supervised machine learning approaches still heavily rely on a large amount of annotated data for training, and thus have limited ability to generalize to these complex data in real-world applications. To overcome this limitation, my long-term research goal is to build **robust intelligent systems that evolve with changes in the environment and interact with people speaking different languages**. With this overall goal in mind, I have focused on addressing the limitations of current NLP systems detailed below.

Background: Over the past decade, the phenomenal success of NLP systems has been mostly driven by supervised learning approaches on a large amount of labeled data. However, supervised machine learning approaches usually impose an ideal assumption that the training and test data both come from a single source. This assumption usually fails in practice, especially when NLP systems are deployed to deal with real text data coming from diverse domains (e.g., social media or medical articles), written in different languages (e.g., Chinese or Swahili), or even associated with different data modalities (e.g., images or structural data). As a result, these data discrepancy issues at the training and testing stages challenge the generalization ability of supervised learning systems in many NLP applications over diverse domains, languages, and modalities. For example, in highly sensitive scenarios such as automatic reply in social media, or translation of news reports, if a model is trained on a biased or out-of-distribution dataset that does not accurately represent the real use case, then deploying this model directly to the test stage without any adaptations could result in skewed predictions, and even raise broader ethical issues, such as translating a third-person pronoun as "he" rather than "she" all the time, or generating racist tweets by Microsoft's chatbot Tay[1]. Hence, I believe one of the key challenges for building the next-generation AI systems in the coming years is the development of **robust AIs adaptive to dynamic changes in their environment**.

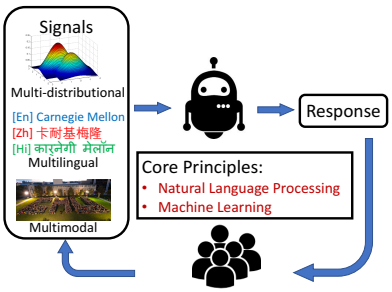

Figure 1: NLP/ML techniques for Diverse Signals in Multilingual Human-Machine Communication

Key Contributions: Within the broad area of artificial intelligence and natural language processing, my Ph.D. research is concerned with the aforementioned data discrepancy issues in three threads:

1. The first major thread of my work focuses on advances of **model robustness in core NLP technologies** (§1), which enables us to analyze, translate or respond to textual inputs from diverse topical domains or distributions.

2. The second thread of my research focuses on **multilingual models for cross-lingual generalization** (§2). This thread extends the core NLP advances to handle over 7000 languages in the world and democratizes NLP techniques to serve a broader population.

3. The third thread of my research focuses on **learning models of natural language associated with other modalities** (§3), because nature language, which exists in abundance as abstract communication symbols, is often complemented by surrounding context in other modalities.

1. `https://en.wikipedia.org/wiki/Tay_(bot)`

1. Model Robustness in Core NLP Technologies

I have devoted a large amount of my research effort to advances in core NLP technologies, because research on natural language stands on the shoulder of computational technologies due to the intrinsic flexibility of natural language. My work has covered a wide range of areas, from lower-level tasks such as sequence tagging [11, 17] and syntactic tree encoding [8] to higher-level applications such as question answering [20, 21] or dialog [19]. Among these tasks, I have worked most extensively on the robustness of machine translation, i.e., the technology to translate between two human languages.

One example of my work in this area focuses on properly translating sentences from very different topical domains, such as medical articles or religious books. When translating textual data between languages, machine translation models crucially rely on the data used to train them. However, it is often expensive and sometimes infeasible to collect labeled data covering all possible topical domains. As a result, the distributional shift between the training and test data is ubiquitous, and neural network based machine translation models may fail dramatically to translate words in the out-of-distribution context, resulting in distorted outputs in practice [7]. In a series of studies, I have assessed and remedied the distributional shift in neural machine translation, and proposed unsupervised adaptation methods for further improving translation accuracy of out-of-distribution sentences based on creating pseudo labeled data by a translation dictionary induced from unaligned monolingual data [18], or enhancing neural machine translation models with domain features [4, 5]. I have also proposed methods that are capable of actively requesting professional translators to annotate a compact set of out-of-distribution unlabeled sentences and phrases in a loop [15], allowing neural machine translation models to be continuously trained on human feedback. These methods have significantly advanced the robustness of neural machine translation models in the unsupervised adaptation setting for which no parallel data in the same domain as the test data exists for training.

In addition, I have proposed methods to estimate translation quality of machine translation outputs by contextual information in the sentences, achieving the first place in three of the six tracks for word-level quality estimation at WMT18 [11]. Besides, in work performed with collaborators at CMU, we developed `compare-mt` for comparing the outputs of multiple translation systems [9], and also won the best demo nomination at NAACL 2019.

2. Cross-lingual Generalization for Natural Language Understanding

The second thread of my work investigates the generalization of core NLP techniques across languages by both empirical [17] and theoretical analysis [22], extending recent NLP advances to a broader application on non-English languages. In NLP research, a vast majority of studies are conducted on English, however, 95% of the world's population does not speak English as their first language, and 75% of the world's population does not speak English at all.[2] At the same time, building NLP models for many languages is extremely challenging largely due to a stark lack of data. Fortunately, many languages share similar syntax or vocabulary, and thus multilingual learning approaches [3, 2] that train models on multiple languages while leveraging this shared structure have begun to show promise as ways to alleviate data sparsity. However, evaluations of these multilingual models are often performed on a very limited and often disparate set of tasks and on typologically similar languages. In my recent work [17], I proposed the Cross-lingual TRansfer Evaluation of Multilingual Encoders (XTREME) benchmark, which covers 40 typologically diverse languages spanning 12 language families, and includes 9 tasks that require reasoning about different levels of syntax or semantics. Together with my collaborators at Google and Deepmind, we released an online platform for the evaluation, and I also provided a set of strong baseline models on all tasks with open-sourced code to facilitate adoption. This benchmark has seen significant uptake in the research community, with 67 papers citing or using the resource in their research since the release in March 2020. In follow-up work, I also proposed a new method for

2. `https://en.wikipedia.org/wiki/List_of_languages_by_total_number_of_speakers`

learning multilingual encoders on parallel data using two explicit alignment objectives that align the multilingual representations at the sentence and word levels [14]. Our cross-lingual transfer results on XTREME show that multilingual models can be trained more efficiently on parallel data using my proposed alignment method, which reduces the heavy reliance on computational resources.

In addition, collaborating with researchers at CMU, we applied multilingual training to neural machine translation [1, 10] to provide rapid response to natural disasters in the DARPA-sponsored Low Resource Human Language Technologies (LORELEI) program. I was the key member contributing to the machine translation task, and we achieved the highest BLEU score in three out of four language-checkpoint pairs in the MT evaluation in 2018.

3. Multimodal Machine Learning for Language and Vision

The final thread of my research focuses on learning NLP models on text that is not isolated from the world around it, but is associated with various signals in other modalities such as visual or audio information. Natural language often serves as an efficient way to map symbols to real-world experiences, while the information in other modalities such as visual or vocal signals helps us to interpret the real-world experiences in other sensory aspects. As a result, for intelligent agents to comprehensively understand the world around us, they need to be able to process and associate input signals from multiple modalities. One example in this area is my work on visual storytelling [12], the task of understanding a stream of pictures and generating a sequence of sentences to describe a coherent story. Most existing visual storytelling methods focus on optimizing standard automatic metrics, which do not necessarily optimize the quality of the story. In contrast, I investigated this problem from a different angle by looking deep into what defines a natural and topically-coherent story, and proposed three assessment criteria which I observe through empirical analysis could constitute a "high quality" story to the human eye. I further proposed a reinforcement learning framework, with reward functions designed to capture the essence of these three quality criteria. Moreover, information from natural language and vision inherently differs, and even human-written sentences cannot cover the full content of an image. Therefore in work with my collaborators at Microsoft and UIUC, we also proposed a fine-grained evaluation method [6], which evaluates generated sentences not only based on how well the machine-generated sentences match the human-written references, but also on how well a caption reflects the information of the image content.

On the other hand, natural language is not only used to describe the world around us, but also used in the interaction between humans and machines. My work on visually grounded question answering [13], for example, has proposed an attention-based model that is trained to answer a textual question by pointing to a correct object in an image. Collaborating with robotic researchers at CMU, I also proposed methods to teach an agent how to understand natural language instructions to navigate in a simulated indoor environment [16].

4. Future Research

My prior research has highlighted the importance of robust NLP systems that process diverse signals for natural language understanding and generation. Moving forward, my mission is to build a robust AI framework that unifies the cognitive abilities of natural language understanding and generation to break down the barriers over all varieties of human-machine communication. With this philosophy in mind, I plan on pushing forward in the following research areas:

Robust Continual Machine Learning for Interactive Communication: With the development of the Internet and mobile, astonishingly more than one quintillion bytes of data are created without labels by human every day, and thus it is infeasible to annotate all emerging data, which leads to outdated supervised learning models vulnerable to out-of-distribution queries or even attacks. Moreover, there

has also been a tendency to train and test machine learning models in a collected static dataset that cannot appropriately represent the dynamic nature of real-world environment. Going forward, first I plan to expand the research to a continual learning paradigm, investigating how machine learning models can continuously learn from dynamic out-of-distribution unlabeled data in the human-machine interaction. One of my recent work on active learning [15], for example, is a preliminary step towards this direction which trains a machine translation model to continuously learn from human feedback. Besides, I also look forward to future opportunity to collaborate with researchers in the Human-Computer-Interaction (HCI) community to incorporate sociology theories in this continual learning paradigm. Secondly, I am also interested in meta-learning algorithms and self-supervised pre-training methods that provide a robust initialization of neural network models and allow few-shot learning from a handful examples in the interaction. Finally, I plan to investigate robust algorithms to mitigate bias present in the training data during the interaction, preventing the models from learning harmful messages such as racism and sexism in real-world applications, and pursuing fairness in NLP technologies.

Multilingual NLP for a Broader Population: Many existing NLP systems have often been built to search and respond in the same language as the human queries. However, training separate NLP systems for each language remarkably increases the burden of maintaining all these systems, and makes it difficult to share knowledge across languages. My prior work on cross-lingual evaluation [17] and alignment [14] has shown great potential of universal representations across languages. Over the next years, I am most excited about a single interactive system that can respond to human speaking different languages, and further bridge the gap between human-human and human-machine communication more efficiently. To realize this target, first I plan to focus on interlingual representation learning by disentangling the language-agnostic knowledge that shares among languages, and language-specific knowledge that uniquely identifies languages themselves. The future studies include the morphological analysis of word tokenizations, the architecture design of multilingual encoding models, linguistically-motivated transfer learning algorithms, and collections of valuable multilingual corpus. Secondly, I would like to apply multilingual NLP techniques to a wide range of real-world applications in order to serve a broader population. One of my on-going work funded by Google, for instance, is aiming to apply multilingual machine translation models to translate Wikipedia articles from English to many other languages, and encourage bilingual speakers to provide post-editing feedback to improve the translation performance. After the procedure of machine translation and human post-edition, we aim to create multilingual Wikipedia articles in high quality for a wide range of users.

Knowledge-based Language Learning for Factual Responses: In much of my previous work regarding multimodal machine learning, I have mainly focused on training neural network models from unstructured data such as textual and visual data. However, with a demand of deeper semantics in human-machine communication, I believe that one key missing component of multimodal learning is the utilization of structured knowledge bases such as WikiData, Freebase, etc. As a result, a final element of my near future research agenda is to pursue methods to integrate explicit knowledge bases with highly parameterized machine learning models for grounded language learning. First, I am interested in the development of efficient encoding methods for structural data and controllable language generation from structural information. My on-going work, for instance, is investigating methods to perform pre-training of multilingual neural models with WikiData knowledge base for downstream sentence generation tasks, such as machine translation and document summarization. Secondly, I also plan to investigate methods to perform knowledge-base language learning on factual data. This includes the analysis on the knowledge captured by pre-trained models, and the design of inference algorithms.

In conclusion, these future directions are both fascinating and challenging, and I believe that the work along these directions will take us a significant step closer to the rich, amazing experience of **multilingual human-machine communication** in our daily lives, just as Internets revolutionized the way we communicate and share information.

References

[1] Aditi Chaudhary, Siddharth Dalmia, Junjie Hu, Xinjian Li, Austin Matthews, Aldrian Obaja Muis, Naoki Otani, Shruti Rijhwani, Zaid Sheikh, Nidhi Vyas, et al. The ariel-cmu systems for lorehlt18. *arXiv:1902.08899*, 2019.

[2] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8440–8451, July 2020.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 4171–4186, June 2019.

[4] Zi-Yi Dou, **Junjie Hu**, Antonios Anastasopoulos, and Graham Neubig. Unsupervised domain adaptation for neural machine translation with domain-aware feature embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1417–1422, November 2019.

[5] Zi-Yi Dou, Xinyi Wang, **Junjie Hu**, and Graham Neubig. Domain differential adaptation for neural machine translation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation (WMT)*, pages 59–69, November 2019.

[6] Ming Jiang, **Junjie Hu**, Qiuyuan Huang, Lei Zhang, Jana Diesner, and Jianfeng Gao. REO-relevance, extraness, omission: A fine-grained evaluation for image captioning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1475–1480, November 2019.

[7] Philipp Koehn and Rebecca Knowles. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation (WMT)*, pages 28–39, August 2017.

[8] Rui Liu*, **Junjie Hu***, Wei Wei*, Zi Yang*, and Eric Nyberg. Structural embedding of syntactic trees for machine comprehension. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 815–824, September 2017.

[9] Graham Neubig, Zi-Yi Dou, **Junjie Hu**, Paul Michel, Danish Pruthi, and Xinyi Wang. compare-mt: A tool for holistic comparison of language generation systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL Demonstrations)*, pages 35–41, June 2019.

[10] Graham Neubig and **Junjie Hu**. Rapid adaptation of neural machine translation to new languages. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 875–880, October-November 2018.

[11] **Junjie Hu**, Wei-Cheng Chang, Yuexin Wu, and Graham Neubig. Contextual encoding for translation quality estimation. In *Proceedings of the Third Conference on Machine Translation (WMT)*, pages 788–793, October 2018.

[12] **Junjie Hu**, Yu Cheng, Zhe Gan, Jingjing Liu, Jianfeng Gao, and Graham Neubig. What makes a good story? designing composite rewards for visual storytelling. In *Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI)*, February 2020.

[13] **Junjie Hu**, Desai Fan, Shuxin Yao, and Jean Oh. Answer-aware attention on grounded question answering in images. In *AAAI Fall Symposium on Natural Communication for Human-Robot Collaboration*, 2017.

[14] **Junjie Hu**, Melvin Johnson, Orhan Firat, Aditya Siddhant, and Graham Neubig. Explicit alignment objectives for multilingual bidirectional encoders. *arXiv:2010.07972*, 2020.

[15] **Junjie Hu** and Graham Neubig. Phrase-level active learning for neural machine translation. *preprint*, 2020.

[16] **Junjie Hu**, Jean Oh, and Anatole Gershman. Learning lexical entries for robotic commands using crowdsourcing. In *AAAI Conference on Human Computation (HCOMP)*, 2016.

[17] **Junjie Hu**, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning (ICML)*, July 2020.

[18] **Junjie Hu**, Mengzhou Xia, Graham Neubig, and Jaime Carbonell. Domain adaptation of neural machine translation by lexicon induction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 2989–3001, July 2019.

[19] Liu Yang, **Junjie Hu**, Minghui Qiu, Chen Qu, Jianfeng Gao, W Bruce Croft, Xiaodong Liu, Yelong Shen, and Jingjing Liu. A hybrid retrieval-generation neural conversation model. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM)*, pages 1341–1350, 2019.

[20] Zhilin Yang, Bhuwan Dhingra, Ye Yuan, **Junjie Hu**, William W Cohen, and Ruslan Salakhutdinov. Words or characters? fine-grained gating for reading comprehension. In *International Conference on Learning Representations (ICLR)*, 2016.

[21] Zhilin Yang, **Junjie Hu**, Ruslan Salakhutdinov, and William Cohen. Semi-supervised QA with generative domain-adaptive nets. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1040–1050, July 2017.

[22] Han Zhao, **Junjie Hu**, and Andrej Risteski. On learning language-invariant representations for universal machine translation. In *International Conference on Machine Learning (ICML)*, July 2020.