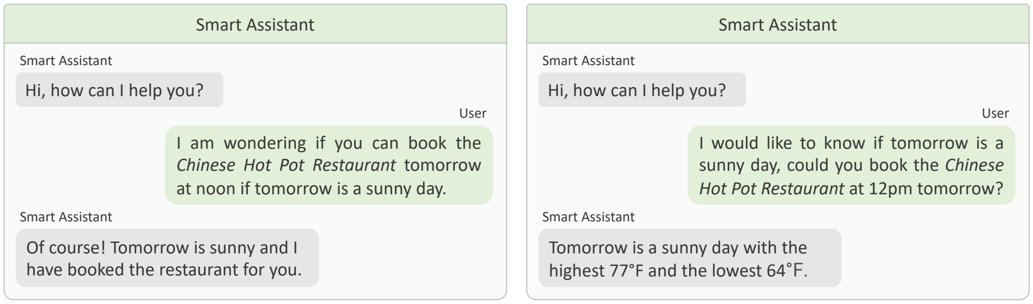
Research Statement

Kuan-Hao Huang, University of California, Los Angeles

Over the past few years, research in artificial intelligence (AI) achieves incredible success. Machines become more and more intelligent and can even outperform humans in some natural language processing (NLP) benchmarks. People start to imagine how NLP applications, such as smart assistants and chatbots, can change our life. Although it looks promising, recent studies have shown that NLP systems are not as reliable and robust as we expect [1, 8]. Sometimes, a small change in the input text can make NLP systems have totally different behaviors, even if the change does not alter the meaning of the text at all. Figure 1 illustrates one example when applying NLP models to a smart assistant system. Users ask the smart assistant the same question in different ways but get very different responses, which deviates from users' expectations. One of the reasons is that existing NLP models are not robust enough to well capture the semantics of texts. Existing NLP models do not connect semantically similar texts well. Hence, a small modification in the input text, such as replacing a word with its synonym or changing the word order, would largely change the behaviors of NLP systems. The robustness issue becomes a big obstacle to making NLP techniques more realistic and hinders people from developing reliable real-world NLP applications. Moreover, this limitation turns more severe for some particular domains, such as the biomedical domain and the public health domain, where we usually do not have large domain-specific text corpora to build NLP models. However, those domains have many important AI applications, such as diagnostic assistants and pandemic predictions. Therefore, how to build reliable and robust NLP systems has become a valuable and indispensable research topic in AI.



Smart Assistant

Smart Assistant

Hi, how can I help you?

User

I am wondering if you can book the *Chinese Hot Pot Restaurant* tomorrow at noon if tomorrow is a sunny day.

Smart Assistant

Of course! Tomorrow is sunny and I have booked the restaurant for you.

Smart Assistant

Smart Assistant

Hi, how can I help you?

User

I would like to know if tomorrow is a sunny day, could you book the *Chinese Hot Pot Restaurant* at 12pm tomorrow?

Smart Assistant

Tomorrow is a sunny day with the highest 77°F and the lowest 64°F.

Figure 1: An example of how NLP systems behave differently for semantically similar texts.

My research focuses on making NLP models more reliable, robust, and realistic. Robust NLP systems should have consistent behaviors for semantically similar texts, even if those texts are written in different word orders, different sentence structures, and even different languages. I believe that improving the NLP model's ability to understand and capture high-level semantics of texts is one of the ways to approach this goal. Once NLP models can rethink and reorganize the meaning and the concepts of a text after reading it, they can have expected behaviors for semantically similar texts. **My prior work focuses on two particular aspects of NLP model robustness: syntax and languages**. Specifically, I study how NLP models behave for semantically similar texts with different syntax (e.g., paraphrase sentences) or different languages (e.g., translation sentences), what factors would make NLP models not robust and therefore have unexpected behaviors, and what techniques can fix the robustness issue. My work has been published in several top-tier conferences in the NLP area (e.g., ACL, EMNLP, and NAACL) and has been cited more than 330 times, according to Google Scholar. My research benefits various NLP tasks, including text representations [7], paraphrase generation [10, 12], and event extraction [11, 5, 6, 13], which are the foundations to build many real-world AI applications. Some of my proposed techniques have been deployed to real-world applications (e.g., Alexa voice assistant). In the future, I plan to keep working on building reliable and robust NLP models, and therefore reducing the gap between NLP benchmarks and real-world AI applications.

1 Improving NLP Model's Robustness to Syntax

Semantics and syntax are unarguably the two most important components to compose a text. Semantics refers to the meaning of texts. For instance, *"John sends a gift to Mary."* and *"Mary sends a gift to John."* describes two different semantics, although the word compositions of the two sentences are the same. In contrast to semantics, syntax refers to the structure and the rules of grammar. For example, *"This restaurant is pretty good."* has a valid syntax in English while *"Restaurant good is pretty this."* has an invalid syntax in English. In other words, a text consists of semantics and valid syntax.

Unfortunately, we found that the way that current NLP models understand the meaning of texts strongly relies on syntax. For instance, the current NLP systems would think the meaning of the following two sentences is quite different: *"We will go hiking if tomorrow is a sunny day."* and *"If it is sunny tomorrow, we will go hiking."*, since they have very different word orders and sentence structures. Another example is *"We will go hiking if tomorrow is a sunny day."* and *"We will go swimming if tomorrow is a sunny day."*, where the NLP system may consider they have similar semantics because there is only one word differing from each other. The sensitivity to the syntax makes the NLP models sometimes have different behaviors for semantically similar texts with different syntax.

To overcome this challenge, I developed several techniques to improve the NLP model robustness in terms of syntax [10, 7, 12]. I proposed a framework to *disentangle* the semantics and the syntax of a text. As shown in Figure 2, the information of a text will be encoded to two separate embeddings: semantic embedding and syntactic embedding. The semantic embedding preserves only the semantic information of the text and no syntactic information. Similarly, the syntactic embedding contains only the syntactic information of the text and no semantic information. The disentanglement of semantics and syntax encourages NLP models to extract the semantics of texts without being affected by the syntax too much. Therefore, no matter how syntax changes, NLP models can always capture the same meaning for semantically texts and become more robust.

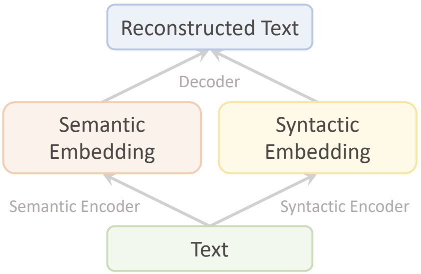


Figure 2: Disentangling the semantics and the syntax of a text.

I studied different ways to learn such disentanglement according to the different types of training data we can access. I demonstrated that the disentanglement can be learned from a large amount of *unannotated texts* [10], can be improved with *annotated* paraphrases pairs [7], and can be further improved with more *fine-grained annotations* such as abstract meaning representations [12]. This leads to more robust NLP models that benefit different NLP downstream tasks, including sentence similarity measurement, paraphrase generation, and few-shot text classification.

2 Improving NLP Model's Robustness to Languages

Supporting multiple languages is one of the important features of real-world NLP applications. Different languages usually mean different vocabularies and different grammar rules, which can be viewed as a much more complicated version of syntactic changes. In recent years, people have shown that NLP models can transfer the learned knowledge from language to language [2]. The underlying factor to makes this cross-lingual transfer successful is the model's robustness to languages, where NLP models can behave similarly for texts with similar meanings but written in different languages. I studied the NLP model's robustness in terms of languages by considering the *zero-shot cross-lingual transfer problem*, where the NLP models are trained with some source languages and directly tested on other unseen target languages.

I pointed out that one key to improving the model's robustness to languages is the *alignment between* *languages*. Although the pre-trained multilingual language models [3, 2] have shown some alignments, they are not perfect enough. My previous work [9] draws connections between the failure cases of zero-shot cross-lingual transfer and adversarial examples [4] and proposes several techniques to make NLP models *robust to noisy alignments between languages*. This makes NLP models transfer knowledge between languages better and leads to robust performance on several zero-shot cross-lingual text classification tasks.

I also explored different ways to improve the alignment between languages. In my prior work [11, 14], we create a *language-agnostic space* to store task-specific knowledge. Because this space is independent of languages, NLP models can understand texts and behave independently of languages as well. We demonstrated improving the model's robustness to languages indeed leads to performance improvements for various NLP tasks, including event argument extraction [11], intent classification [14], and slot filling [14].

3 Future Research Directions

My long-term goal is to make NLP models reliable, robust, and realistic enough to build trustworthy real-world NLP systems. I have the following research plan to achieve this goal from different perspectives.

Making NLP models understand texts like humans. My prior work [10, 7, 12] has shown that improving the model's ability to understand texts and extract concepts is the key to robustness to syntactic changes. I plan to keep exploring this direction as follows.

- **Learning better text representations to capture semantics.** Most current approaches for learning sentence representations consider only paraphrase pairs and entailment pairs as the training data. Although paraphrase pairs and entailment pairs provide very strong semantic relations between texts, they usually cover limited domains and are not general enough. Therefore, the learned text representations might not perform well for particular domains (e.g., biomedical domain). On the other hand, there exist tons of NLP tasks across various domains that consist of the texts and the corresponding label strings. Although those texts and label strings are not semantically the same, they actually describe some *weak semantic relations* between words and sentences. I will study how to leverage those weakly supervised signals to learn text representations that capture semantics better.
- **Concept-based text representations.** Existing NLP models usually encode texts into a series of vectors that are very hard to explain and interpret. I believe *concept-based text representations* are a better way for NLP models to understand texts. When humans read texts, we understand texts by extracting a lot of concepts, such as subject, object, main verb, topic, tense, and other abstract meanings. Therefore, I would like to study how to encode texts into *a list of concepts*. We can create concepts by leveraging existing NLP tasks, such as sentiment classification, topic classification, and relation extraction. As more and more concepts are included, NLP models with concept-based text representations can understand texts more like humans and become explainable and interpretable.
- **Benchmarking new semantic measurements.** I realized that most current NLP benchmarks for measuring semantics capturing are not good enough to measure *fine-grained semantics capturing*. They only consider a single view to understand texts while human usually has multiple views to understand texts. For example, if we care about the time, the semantic similarity between *"I reserved a Chinese restaurant at 12pm tomorrow."* and *"I reserved a Chinese restaurant tomorrow night."* should be higher than the semantic similarity between *"I reserved a Chinese restaurant at 12pm tomorrow."* and *"I reserved a French restaurant at 12pm tomorrow."* In contrast, if we care about the restaurant more, the semantic similarity of those sentence pairs will change accordingly. Current NLP models usually encode a text to a fixed representation and therefore they do not support *multi-view understanding*. How to design models to learn such *conditional text representations* is a novel, interesting, and valuable research direction.

Robust multilingual NLP systems. I will keep working on improving NLP models' robustness to languages since it benefits many NLP applications. The following are some directions that I am particularly interested in.

- **Learning better alignment between languages.** As pointed out by my prior work [9, 11], the alignment between languages is one of the keys to improving the model's robustness to languages. Unlike most previous studies that consider translation word pairs or translation sentence pairs to further align pretrained multilingual representations, I plan to consider syntactic features, such as dependency parses and constituency parses, to extract the shared knowledge from different languages. Those syntactic features have the potential to capture the syntactic similarity between languages and therefore learn better alignment between languages.
- **Handling code-switching texts.** Another interesting direction I would like to focus on is code-switching texts. Most existing work usually assumes the input text contains only one language. When code-switching texts or mixed-language texts are presented, the existing NLP system might be failed. However, code-switching texts or mixed-language texts are common for some applications, such as smart assistants and diagnostic assistants, since there are always some terms that cannot be translated. I plan to start by collecting large-scale code-switching data by either crowd-sourcing or data synthesis. Then, I will study how to extend existing techniques to improve the model's robustness to code-switching texts.

Trustworthy NLP system against malicious behaviors. For real-world NLP applications, sometimes we will face malicious behaviors that try to hurt or attack the system. How to prevent and protect the system from malicious attacks is one of the most important things to building a trustworthy system.

- **Discovering malicious behaviors.** There have been several studies showing that we can intentionally design malicious input texts that make NLP models behave abnormally, called adversarial attacks [1, 8]. How to detect and discover those malicious examples is an important research direction for improving the robustness of NLP models. Unlike most existing work that focuses on character-level or word-level attacks, I will put more attention on sentence-level attacks, such as paraphrasing and sentence structure modifications, because they are more similar to real-world situations.
- **Defending adversarial attacks.** Since most adversarial attacks are about syntactic modifications of texts, my work studying syntactic robustness [10, 12] can be a good starting point for defending adversarial attacks. I plan to extend my work to build effective robust models against malicious behaviors. Some possible directions are generating diverse syntactic training examples and fixing spurious correlations.

Grounding NLP techniques to real-world AI applications in different areas. My prior work has shown improvements for many NLP benchmarks, including text representations [7], paraphrase generation [10, 12], and event extraction [11, 5, 6, 13]. I believe those techniques can be the foundation to build real-world AI applications beyond the NLP area. I look forward to collaborating with experts in different areas. Particularly, I am interested in collaborating with experts in biomedical domains to build robust diagnostic assistants and collaborating with experts in public health domains to build pandemic detectors or predictors based on the texts from social media. I believe that my research can make NLP models more reliable, robust, and realistic, and therefore build many valuable AI applications.

References

[1] Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani B. Srivastava, and Kai-Wei Chang. Generating natural language adversarial examples. In *EMNLP*, 2018.

[2] Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. In *NeurIPS*, 2019.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019.

[4] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015.

[5] I-Hung Hsu*, **Kuan-Hao Huang***, Elizabeth Boschee, Scott Miller, Prem Natarajan, Kai-Wei Chang, and Nanyun Peng. DEGREE: A data-efficient generation-based event extraction model. In *NAACL*, 2022. (*equal contribution).

[6] I-Hung Hsu*, **Kuan-Hao Huang***, Shuning Zhang, Wenxin Cheng, Premkumar Natarajan, Kai-Wei Chang, and Nanyun Peng. A simple and unified tagging model with priming for relational structure predictions. *arXiv preprint arXiv:2205.12585*, 2022. (*equal contribution).

[7] James Y. Huang, **Kuan-Hao Huang**, and Kai-Wei Chang. Disentangling semantics and syntax in sentence embeddings with pre-trained language models. In *NAACL*, 2021.

[8] Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. Adversarial example generation with syntactically controlled paraphrase networks. In *NAACL*, 2018.

[9] **Kuan-Hao Huang**, Wasi Uddin Ahmad, Nanyun Peng, and Kai-Wei Chang. Improving zero-shot cross-lingual transfer learning via robust training. In *EMNLP*, 2021.

[10] **Kuan-Hao Huang** and Kai-Wei Chang. Generating syntactically controlled paraphrases without using annotated parallel pairs. In *EACL*, 2021.

[11] **Kuan-Hao Huang***, I-Hung Hsu*, Prem Natarajan, Kai-Wei Chang, and Nanyun Peng. Multilingual generative language models for zero-shot cross-lingual event argument extraction. In *ACL*, 2022. (*equal contribution).

[12] **Kuan-Hao Huang***, Varun Iyer*, Anoop Kumar, Sriram Venkatapathy, Kai-Wei Chang, and Aram Galstyan. Unsupervised syntactically controlled paraphrase generation with abstract meaning representations. In *EMNLP-Findings*, 2022. (*equal contribution).

[13] Tanmay Parekh, I-Hung Hsu, **Kuan-Hao Huang**, Kai-Wei Chang, and Nanyun Peng. Geneva: Pushing the limit of generalizability for event argument extraction with 100+ event types. *arXiv preprint arXiv:2205.12505*, 2022.

[14] Fei Wang, **Kuan-Hao Huang**, Anoop Kumar, Aram Galstyan, Greg Ver Steeg, and Kai-Wei Chang. Zero-shot cross-lingual sequence tagging as seq2seq generation for joint intent classification and slot filling. In *MMNLU Workshop@EMNLP*, 2022.