Research Statement
Mingda Chen

How can we build computational models that are capable of understanding and learning from the traces left by humans? To tackle this problem, my research seeks to tailor naturally-occurring textual data to serve as training corpora for various purposes and to establish challenging evaluation tasks. Written language is an ubiquitous tool for communication and documentation. There are plenty of freely-available textual data: e.g., Wikipedia,\(^1\) Reddit,\(^2\) and Fandom\(^3\) among others. These data often contain sophisticated knowledge expressed with complex language structures. For example, people base their writings on common ground in knowledge without explicitly describing it (e.g., animal is a more generic category than dog and cat); for encyclopedia, there usually are dedicated annotations for connecting pieces of information scattered around different places for convenience of readers (e.g., hyperlinks that point the same person or events mentioned in different documents to the same place for disambiguating entities). When training neural models, the knowledge and structure are often stripped away from the input due to limited context sizes. I believe incorporating the extra information into the training and evaluation of neural models can help models understand human-written language and solve real-world tasks.

Below I describe some first steps oriented towards these goals.

Characterizing Textual Resources with Properties of Natural Language

Increasing attention has been devoted to the large amount of freely-available textual data. One example is language model pretraining (i.e., word prediction based on nearby context) where recent work has shown the potential value of plain text. However, it is still unclear as to what we can accomplish with these data beyond the word prediction objective. For example, when people write articles, they tend to organize the whole document in a coherent manner. The coherence of a whole document is often neglected if models are only taught to recover words from local context. Besides coherence, there are many other properties in written language, such as coreference and commonsense knowledge among others. These phenomena naturally emerge as a way to make the conveyed information concise and clear.

My work strives to characterize the naturally-occurring textual data using properties of natural language. In particular, I have shown that the rich annotations on Wikipedia are related to discourse (Chen et al., 2019b) (e.g., see Figure 1a), entity representations (Chen et al., 2019a), and entailment relationships for sentence/phrase pairs (Chen et al., 2020a). Some of these work were later incorporated in the large-scale language model pretraining for generic representations (Lan et al., 2020) and in-context few-shot learning (Chen et al., 2021b). Our empirical findings include (1) sentence ordering prediction consistently improves model performances of the pretrained language models on many downstream tasks (e.g., question answering); and (2) the Wikipedia article category hierarchy helps the task of textual entailment classification.

Disentangling Latent Factors for Interpretability and Controllability

Beyond the properties of natural language, I am interested in studying latent factors that affect the observed data. Language is complicated. Its semantic meaning can be affected by various factors, such as intention and

---


\(^2\)[https://www.reddit.com/](https://www.reddit.com/), an online forum for discussion and web content rating.

\(^3\)[https://www.fandom.com/](https://www.fandom.com/), a fan-contributed encyclopedia of movies, films, and other media.
context, with accordingly adjusted discourse and syntactic structures for clarity of expressions. As neural models are notorious for conflating different factors, in this line of work, I seek to disentangle these factors to achieve better interpretability and controllability.

To this end, I have designed a framework for parallel corpora to disentangle syntax and semantics in sentence representations (Chen et al., 2019d). I later extended the framework to the novel task of syntactic exemplar-controllable paraphrase generation (Chen et al., 2019c, 2020b) (e.g., see Figure 1b). The setting is analogous to style transfer in computer vision, in which an image is generated that combines the content from one image and the style from another.

Tailoring Textual Resources for Challenging Evaluation Tasks

I have also found naturally-occurring textual data broadly useful for the tasks that require complex reasoning. In this line of work, I seek to automatically construct datasets from the naturally-occurring textual data in challenging domains (e.g., long-form texts in the domain of television series).

In particular, I have built datasets from Wikipedia that require generating a coherent passage by connecting all the entities in the given tabular data, and the story also needs to fit the provided background knowledge (Chen et al., 2021c). Based on other community-contributed websites, I have built a dataset for TV show transcript summarization that requires drawing information from utterances across a wide range of the input and integrating the information to form concise plot descriptions (Chen et al., 2021a), and a story generation dataset with character descriptions as constraints that requires drawing relevant information from lengthy provided documents about characters based on a brief summary (Chen and Gimpel, 2021).

While recent work in the natural language processing (NLP) community has shown parity with humans on certain NLP or artificial intelligence tasks, initial results on these new tasks finds them to be extremely difficult for current state-of-the-art methods.

Future Directions

Disentangling Latent Factors for Diverse Tasks. I have devised neural models for improving interpretability and controllability using implicit yet natural supervision from paraphrases and machine translations. These approaches can be generalized to any resources that are formed by data pairs, such as dialogues or summarization. They can be used for disentangling the factors that are shared between pairs of inputs and those that are not shared, such as intentions and the personalized styles in dialogues, sentence-level fluency and document-level discourse in sentence modeling, or important events and irrelevant details in summarization. This direction is appealing in that although large pretrained models have yielded superhuman performance, researchers still lack understanding of the behaviors of these models. Better interpretability or controllability could help us improve their robustness and worst-case behaviors so that they can be better applied in user-facing applications.

Commonsense Knowledge for Pretrained Language Models. Another exciting research direction is to analyze/incorporate commonsense knowledge in pretrained language models. Specifically, I am interested in learning domain-specific commonsense from dialogues in certain subreddits\(^4\) (e.g., technical or social) or distilling commonsense knowledge from existing pretrained models. When deploying these models in real-life applications, having such knowledge can make them more reliable.

Unifying Various Types of Learning Signals. Although we have designed modeling choices to take entities and discourse into account, it is still unclear as to what is the best design for a unified model that can incorporate all of these learning signals. In addition, I am also interested in combining discourse, linking, and paraphrase objectives with BERT-like models, as well as other types of natural supervision, such as naturally-occurring bold/italics/underlining annotations in web text, web images/videos/audio paired with text, and long-distance discourse cues like two paragraphs in two chapters, among others.

---

\(^4\)Subreddits are on the social media website Reddit and they are dedicated to a particular topic that people write about.
References


