Lightweight Methods for Simulating Counterfactual Reasoning

Name: Michael Van Ballescas Supranes

Laboratory's name: Social Computing Laboratory

Supervisor's name: Prof. Eiji Aramaki

Abstract:

Counterfactual reasoning is the process of imagining how an observed state or outcome would change if relevant conditions or assumptions were altered. In natural language processing, counterfactual reasoning is useful for developing classifiers that are more robust to spuriously correlated features, commonly through counterfactual data augmentation (CDA). It is also useful for guiding generation when different assumptions should lead to different appropriate outputs. This dissertation investigates lightweight methods for simulating counterfactual reasoning in large language models (LLMs) across three studies. The first study addresses outcome-oriented counterfactual text generation for hate speech classification under limited-resource constraints. In this setting, task-specific counterfactual data may be unavailable, or full fine-tuning of a language model may be impractical. To address this problem, this study proposes *Gradient-assisted Energy-based Sampling* (GENES), a plug-and-play counterfactual text generation method for CDA. The second study examines a limitation of outcome-oriented counterfactual generation, and compares counterfactual-based methods with other data-level interventions. In symptom detection from social media text, outcome labels depend not only on the symptom mentioned but also on related concepts such as evidence type, timing, recovery, and uncertainty. To model these dependencies, this study proposes the *Concept-Aware Learning Framework* (CALF), which leverages few-shot learning and combines multi-attribute counterfactual data augmentation with multi-task learning. It also introduces MedWeb Checklist, a counterfactual-based evaluation dataset for measuring model sensitivity to relevant attributes. Performance on the checklist dataset was found to be a possible indicator of out-of-distribution performance. The third study extends counterfactual reasoning to research idea generation. It first proposes *Trend-aware and Cluster-based Evaluation* (TrACE), a training-free framework for assessing forecastable novelty without relying on LLMs as judges. It then proposes *Steering along Alternative Reasoning Traces* (StART), a training-free method for simulating counterfactual reasoning through activation steering. In this approach, counterfactual reasoning is simulated by sampling contrasting reasoning paths, and activation steering is used to guide LLMs toward the path associated with more novel ideas.