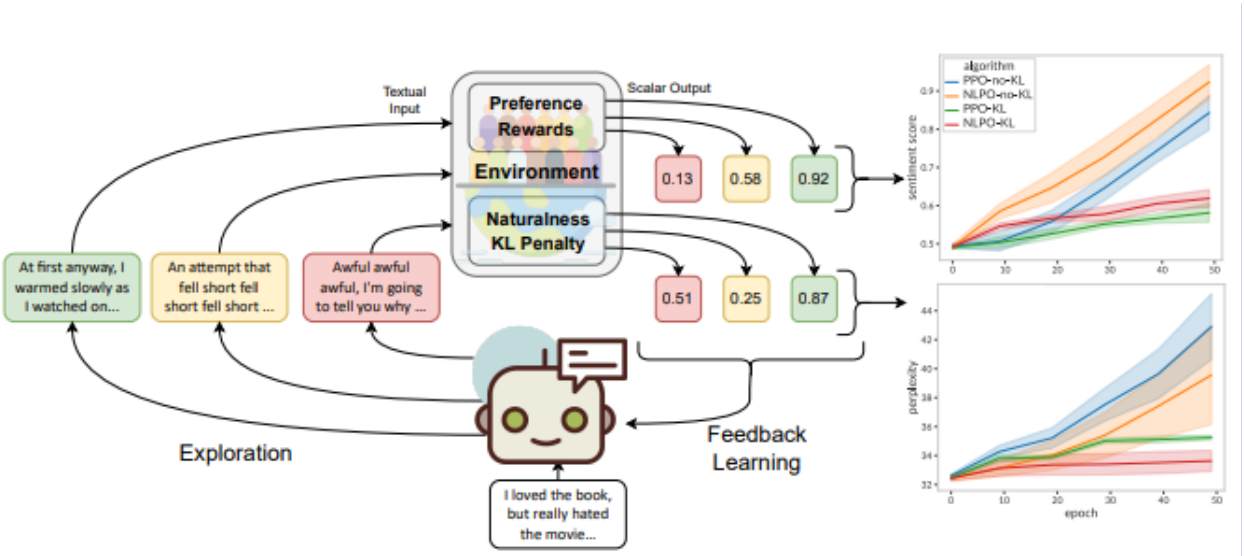
UNIVERSIDAD MAYOR REAL Y PONTIFICIA DE SAN FRANCISCO XAVIER DE CHUQUISACA

TRABAJO DE INVESTIGACION PRACTICA # 2

Universitario (a): Duran Daza José Ignacio Carrera: ING. CIENCIAS DE LA COMPUTACION

Docente: ING. Walter Pacheco

Materia: SIS420



Si vemos la generación de texto como un problema de toma de decisiones secuencial, el aprendizaje por refuerzo (RL) parece ser un marco conceptual natural. Sin embargo, el uso de RL para la generación basada en LM enfrenta desafíos empíricos, incluida la inestabilidad del entrenamiento debido al espacio de acción combinatorio, así como la falta de bibliotecas de código abierto y puntos de referencia personalizados para la alineación de LM. Así, surge una pregunta en la comunidad investigadora: ¿es la RL un paradigma práctico para la PNL?

Para ayudar a responder esto, primero presentamos una biblioteca modular de código abierto, RL4LMs (aprendizaje reforzado para modelos de lenguaje), para optimizar los generadores de lenguaje con RL. La biblioteca consta de algoritmos RL en política que se pueden usar para entrenar cualquier codificador o codificador-decodificador LM en la biblioteca HuggingFace (Wolf et al. 2020) con una función de recompensa arbitraria. A continuación, presentamos el punto de referencia GRUE (Evaluación general de comprensión del lenguaje reforzado), un conjunto de 6 tareas de generación de lenguaje que no son supervisadas por cadenas objetivo, sino por funciones de recompensa que capturan medidas automatizadas de preferencia humana. GRUE es el primer estilo de tabla de clasificación evaluación de algoritmos RL para tareas de PNL. Finalmente, presentamos un algoritmo de RL eficaz y fácil de usar, NLPO (Optimización de políticas de lenguaje natural)} que aprende a reducir efectivamente el espacio de acción combinatoria en la generación de lenguaje. Mostramos 1) que las técnicas de RL son generalmente mejores que los métodos supervisados para alinear los LM con las preferencias humanas; y 2) que NLPO exhibe una mayor estabilidad y rendimiento que los métodos de gradiente de políticas anteriores (p. ej., PPO (Schulman et al. 2017)), según la evaluación automática y humana.