

PERSONALIZED DATA-TO-TEXT NEURAL GENERATION

Apprentissage par renforcement et modèles de langue

20 juillet 2023

Ben KABONGO

Stage - M1 DAC - Sorbonne Université

Sommaire

- Apprentissage par renforcement
- Papier **Is reinforcement learning (not) for natural language processing** Ramamurthy et al. 2022
- Application et discussions : data-to-text personnalisé et RL

Apprentissage par renforcement

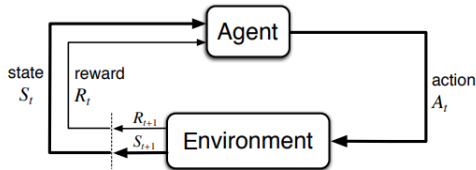


Figure 1 – Apprentissage par renforcement

- Agent, environnement, état, action, récompense
- **Processus de décision markovien** : (S, A, T, R) : ensembles des états et des actions, fonctions de transition et de récompense
- **Politique** : fonction de choix d'une action étant donné un état
- **Objectif** : maximiser la somme des récompenses perçues au cours du temps

Papier

Introduction

Objectifs et problématiques

- Génération de texte = processus de décision markovien
- Aligner les LLMs avec les préférences humaines via le RL
- **Le RL est-il adapté pour le NLP ?** : l'espace des actions est très grand

Apports

- **RL4LMs** : librairie open source pour fine-tuner des LLMs avec du RL
- **GRUE (General Reinforced-language Understanding Evaluation)** : benchmarks pour 6 tâches en NLP abordées avec du RL
- **NLPO (Natural Language Policy Optimization)** : nouvel algorithme de RL pour du NLP

- Librairie open source pour fine-tuner des LLMs avec du RL
- Surcouche de *HuggingFace* (NLP) et *Stable-baselines-3* (RL)
- **Environnement : MDP de génération au niveau du token :**
 - **Action** : token du vocabulaire
 - **Etat** : série de tokens $s_t = (x_0, x_1, \dots, a_0, a_1, \dots)$
 - **Transition** : déterministe : ajout de l'action (du token) à l'état
 - **Récompense** : à la fin de l'épisode
- **Fonctions de récompenses (métriques) :**
 - **n-gram overlap** : ROUGE, BLEU, SacreBLEU, METEOR
 - **model-based semantic** : BertScore, BLEURT
 - **task-specific** : CIDER, SPICE, *PARENT*
 - **diversity, fluency, naturalness** : perplexity, MSSTR, entropie de Shannon, DIST-N
 - **task-specific, model-based human preference** : classification sur les préférences

NLPO : Natural Language Policy Optimization

- Les vocabulaires des LLMs sont très grands : cause de l'inefficacité des méthodes de RL appliquées à du NLP
- **NLPO (Natural Language Policy Optimization)** :
algorithme de RL pour du NLP inspiré de PPO (Proximal Policy Optimization)
<https://r14lms.apps.allenai.org/algorithms>
- NLPO apprend à masquer les tokens moins pertinents dans le contexte au fur et à mesure de l'apprentissage
- Echantillonnage top-p : restreindre le vocabulaire au plus petit ensemble possible de tokens dont la probabilité cumulée est supérieure au paramètre de probabilité p

NLPO : Natural Language Policy Optimization

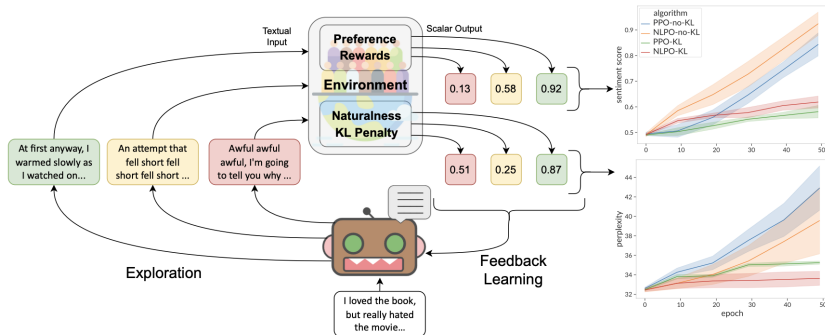


Figure 2 – Natural Language Policy Optimization (NLPO) in the case of sentiment-guided continuation (tiré de Ramamurthy et al. 2022)

GRUE (General Reinforced-language Understanding Eval)

- Collection de 7 tâches de génération NLP : data-to-text, summarization, traduction, etc.
- **Catégories des métriques d'évaluation en test**
 - **Task preference metrics** : mesure de satisfaction des critères de la tâche de génération spécifique
 - **Naturalness metrics** : offrent une perspective sur des facteurs autres que la sémantique : fluidité, de la lisibilité, etc.
- Choix de métriques libre en apprentissage
- **Expérimentations** : comparaisons entre : l'approche supervisée, PPO, NLPO et des approches hybrides
- **Modèles de langue** : GPT-2, T5-base
- **Résultats** : L'approche hybride supervisée + RL (PPO ou NLPO) est meilleure que les autres approches

Data-to-text personnalisé et RL

Problématique et objectifs

- **Data-to-text personnalisé** : du data-to-text où les descriptions textuelles des données sont proches des préférences/styles des utilisateurs
- **Proposition de modèle et de dataset**

Méthodologie

- **Modèle** : fine-tuning d'un LLM (ex : T5) avec du RL
- **Dataset** : déduire un dataset en inférence (sur les données WikiRoto)
- **Choix** : Quelles entrées/sorties ? en apprentissage ? en inférence ? quelles récompenses pour le modèle ?

Données et paramètres

Paramètres du modèle

- **Entrée :**
 - **Données semi-structurées** : ex : infobox des films WikiRoto
 - **Informations sur l'utilisateur** : identifiant, exemple de texte (ex : critique de l'utilisateur sur le film)
- **Sortie** : description textuelle personnalisée pour l'utilisateur des données d'entrée

Rewards

- **Data-to-text/NLP metrics** : PARENT, BLEU, perplexité, similarité avec sortie réelle
- **Personnalisation** :
 - **Sentiment analysis** : on veut une polarité similaire entre l'exemple (la review utilisateur) et la description générée
 - **Authorship attribution** : on veut que la description générée soit attribuée à l'utilisateur

Discussion : Modèles et paramètres

Modèles de reward

- **Modèles pré-entraînés :**
 - Analyse des sentiments : *transformers pipeline sentiment analysis*
 - Authorship attribution : *BertAA*
- **LLMs fine-tuné :** *T5, BERT*
- **Autres méthodes :** *BoW TF-IDF, RNNs, CNNs*

Authorship attribution

- **Nombre d'auteurs :** 5, 10, 20, 50, 80, 100, tous les utilisateurs ? ?
 - 80 utilisateurs ont plus de 2000 reviews => 160 000 exemples
- **Clustering sur les auteurs :** réduction du nombre d'auteurs en regroupant les auteurs semblables (avec du KNN par exemple)

Discussion : Bibliothèques NLP + RL

■ RL4LMs Reinforcement Learning for Language Models :

- **Modèles de langues** : GPT-2, T5

- **Politiques** : NLPO, PPO, Supervisée, hybride

- ++ Reward très personnalisable

- Problèmes de gestion de la mémoire??

■ TRL Transformer Reinforcement Learning :

- **Modèles de langues** : GPT-2, BLOOM, Neo

- **Politiques** : PPO

- Génération de réponse ou de suite sur la base d'une requête qui peut être le début d'une phrase.

- Visiblement pas adapté au data-to-text

- Solution alternative à RL4LMs avec beaucoup d'adaptations

Discussion : TRL

Rollout:



Evaluation:



Optimization:

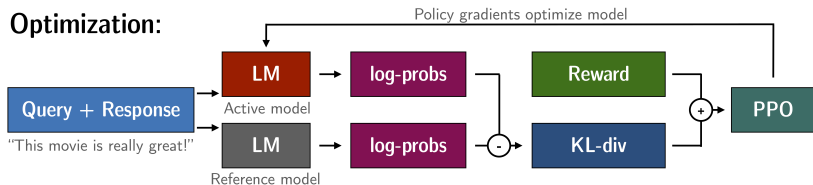



Figure 3 – TRL overview

Références

-  Ramamurthy, Rajkumar et al. (2022). “Is Reinforcement Learning (Not) for Natural Language Processing ? : Benchmarks, Baselines, and Building Blocks for Natural Language Policy Optimization”. In :
url : <https://arxiv.org/abs/2210.01241>.
- **RL4LMs Reinforcement Learning for Language Models :**
 - <https://rl4lms.apps.allenai.org>
 - <https://github.com/allenai/RL4LMs>
- **TRL Transformer Reinforcement Learning :**
 - <https://huggingface.co/docs/trl/index>
 - <https://github.com/lvwerra/trl/tree/main>